Construct Representation of First Certificate in English (FCE) Reading

Michael Corrigan

CONSTRUCT REPRESENTATION OF FIRST CERTIFICATE IN ENGLISH (FCE) READING

by

MICHAEL CORRIGAN

A thesis submitted to the University of Bedfordshire in partial fulfilment of the requirements for the degree of Doctor of Philosophy

January 2015

CONSTRUCT REPRESENTATION OF FIRST CERTIFICATE IN ENGLISH (FCE) READING

by

MICHAEL CORRIGAN

ABSTRACT

The current study investigates the construct representation of the reading component of a B2 level general English test: First Certificate in English (FCE). Construct representation is the relationship between cognitive processes elicited by the test and item difficulty. To facilitate this research, a model of the cognitive process involved in responding to reading test items was defined, drawing together aspects of different models (Embretson & Wetzel, 1987; Khalifa & Weir, 2009; Rouet, 2012). The resulting composite contained four components: the formation of an understanding of item requirements (OP), the location of relevant text in the reading passage (SEARCH), the retrieval of meaning from the relevant text (READ) and the selection of an option for the response (RD). Following this, contextual features predicted by theory to influence the cognitive processes, and hence the difficulty of items, were determined. Over 50 such variables were identified and mapped to each of the cognitive processes in the model. Examples are word frequency in the item stem and options for OP; word frequency in the reading passage for READ; semantic match between stem/option and relevant text in the passage for SEARCH; and dispersal of relevant information in the reading passage for RD. Response data from approximately 10,000 live test candidates were modelled using the Linear Logistic Test Model (LLTM) within a Generalised Linear Mixed Model framework (De Boeck & Wilson, 2004b). The LLTM is based on the Rasch model, for which the probability of success on an item is a function of item difficulty and candidate ability. The holds for LLTM except that item difficulty is decomposed so that the contribution of each source of difficulty (the contextual features mentioned above) is estimated. The main findings of the study included the identification of 26 contextual features which either increased or decreased item difficulty. Of these features, 20 were retained in a final model which explained 75.79% of the variance accounted for by a Rasch model. Among the components specified by

the composite model, OP and READ were found to have the most influence, with RD exhibiting a moderate influence and SEARCH a low influence. Implications for developers of FCE include the need to consider and balance test method effects, and for other developers the additional need to determine whether their tests test features found to be criterial to the target level (such as non-standard word order at B2 level). Researchers wishing to use Khalifa and Weir's (2009) model of reading should modify the stage termed named inferencing and consider adding further stages which define the way in which the goal setter and monitor work and the way in which item responses are selected. Finally, for those researchers interested in adopting a similar approach to that of the current study, careful consideration should be given to the way in which attributes are selected. The aims and scope of the study are of prime importance here.

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate: Signature:

Date:

Table of Contents

List of tables

List of figures

List of equations

Acknowledgements

My thanks goes to all those who helped me to conduct my research and complete this thesis. Most significant were Professor Tony Green, Dr Nick Saville and Dr Francesca Parizzi. Professor Green has been my Director of Studies for the last five years and has been a great help in guiding me through the process of doing a PhD part-time, reading and commenting on many drafts and even advising me on how best to tackle the various administrative hurdles which present themselves from time to time. Dr Saville was my line manager when I began the PhD and for most of the time I have been working on it. Since my PhD was sponsored by my employer, his support was crucial, and his words of advice, on various aspects, were wise. Dr Parizzi, or Francesca, as I usually refer to my partner, was able to offer me much good advice about the topic, assistance with qualitative analysis and the process of doing a PhD, as she works and has studied in the same field. She also gave up many hours we could have spent together and also, at many significant moments, lightened the load life wanted to place on my shoulders and provided me with encouragement to continue my studies.

I also offer my thanks to many others. Dr Muhammad Naveed Khalid, my colleague and supervisor, gave me many useful suggestions about the psychometric methodology I used, as well as giving insightful comments on the text of the thesis. Dr Fumiyo Nakatsuhara, my second supervisor at CRELLA, after reading the first complete draft of the thesis, provided valuable comments on how it might be improved and Professor Liz Hamp-Lyons, although one of my supervisors for a very short time, and therefore, only able to offer me a few words of advice, chose very sage ones.

The data and materials used for this study was supplied by my employer, Cambridge English Language Assessment. I would like to thank those involved in locating and collating what I needed. These include Ron Zeronis and colleagues in Assessment, and Chris Bell and Laurence Calver in Research and Validation.

The study was, at an earlier stage, to be a comparison between two tests: FCE and CELI 3, a test of Italian provided by the Il Centro per la Valutazione e le Certificazioni Linguistiche (CVCL) dell'Università per Stranieri di Perugia. I would therefore like to thank Danilo Rini, Paola Ramaccioni and Professoressa Giuliana Grego Bolli for all the help they gave me and their

good will for the research, which, it should be mentioned, extended to the FCE portion of the study. I am only sorry that I was not able to include CELI 3 in the study.

My thanks also goes to numerous colleagues and those working or studying at CRELLA who gave me help, good advice or simply expressed their good will. Among these are Jane Lloyd, Dr Andrew Somers, Dr Evelina Galaczi, Dr Szilvia Papp, Dr Fiona Barker, Dr Agnieszka Walczak, Dr Angeliki Salamoura, Dr Nahal Khabbazbashi, Mark Elliott, Dr Ardeshir Geranpayeh, Carrie Warren, Fiona Beedle, Becky Bullett, Professor Mike Milanovic, Dr Lynda Taylor, Professor Cyril Weir, Professor Stephen Bax, Dr John Field, Dr Sathena Chan and Professor Roger Hawkey. John Savage deserves a special mention as a very willing and flexible proof reader, who actually wanted to read my thesis. Something I do not need to understand but am very grateful for.

This thesis reports on research using examination data provided by Cambridge English Language Assessment.

List of abbreviations

AIC	Akaike Information Criterion
ASVAB	Armed Services Vocational Aptitude Battery
AWL	Academic Word List
BIC	Bayesian Information Criterion
BNC	British National Corpus
CAE	Certificate in Advanced English
CEFR	Common European Framework of Reference: Learning, Teaching, Assessment
CELEX	Centre for Lexical Information
CFA	Confirmatory Factor Analysis
Chi Df	Chi-square Degrees of Freedom
Chisq	Chi-square Difference
CPIDR	Computerized Propositional Idea Density Rater
CPM	Cognitive Psychometric Model
Cumul. %-age	Cumulative Percentage
Cumul. Freq.	Cumulative Frequency
Df	Degrees of Freedom
EFA	Exploratory Factor Analysis
FCE	First Certificate in English
GLMM	Generalized Linear Mixed Model
GLTM	General Multicomponent Latent Trait Model
GRE	Graduate Record Examination
IRT	Item Response Theory
L1	First Language
LCM	Latent Class Model
LD	Local Item Dependency
LLTM	Linear Logistic Test Model
loglik	Log Likelihood
LRT	Likelihood Ratio Test
LSA	Latent Semantic Analysis
Max	Maximum
Min	Minimum
MLTM	Multicomponent Latent Trait Model
MLTM-D	Multicomponent Latent Trait Model for Diagnosis
NI	Necessary Information
Pr(>Chisq)	Probability of Chi-square Value
RD	Response Decision
REML	Restricted Maximum Likelihood
RWLLTM	Random Weights Linear Logistic Test Model
SD	Standard Deviation
SEM	Standard Error of Measurement
Signif	Significance
Std. Error	Standard Error of the Estimate
TBR	Tree-Based Regression
TLU	Target Language Use

TOEFL	Test of English as a Foreign Language
TRACE	Task-based Relevance Assessment and Content Extraction
VIF	Variable Inflation Index
-ve	Negative
+ve	Positive

1 Introduction

1.1 The context of this research

1.1.1 Test constructs and validity

Generally speaking, the purpose of a test is to measure of a quality of interest. A criterion-referenced test of proficiency in English as a foreign language, for example, should provide test results which summarise what each candidate is able to do with that language. The evidence for such a summary is derived from responses to the items and tasks contained by the test. In standardised educational tests, the elicited responses are usually understood to represent only a sample of what a candidate is able to do, as the range of possible opportunities for language use is vast and it is only practical to tap a limited proportion of them (Bachman, 2007; Kane, 2009). Responses collected are therefore considered to be a representative sample of behaviours drawn from a much larger pool (Messick, 1989). However, results should be generalisable: if a different set of items were chosen, the test results should be the same, or very similar, and users of the results should be able to infer the same conclusions about the candidates.

To facilitate test construction and the interpretation of the results, the notion of a testing *construct* is important. Among other things, the definition of the construct has the function of limiting the domain from which behaviours are elicited and, therefore, enables inferences about the results to be targeted on specific domains (ALTE & Council of Europe, 2011). A test of a foreign language may, for example, aim to test the ability to communicate in that language, or to test knowledge of the language, or to test both. The inferences that can be drawn about what the candidate can do are limited accordingly. Other ways to limit, or more accurately define such a construct include identifying the skills to be included in the test and the specificity of the intended *target language use* (TLU). For example, a test may be quite general, or it may target a TLU of academic study or air traffic control (Bachman & Palmer, 2010; Douglas, 2000). Ideally, a test is designed according to a construct definition and recommendations, supplied by the test provider, concerning inferences based on results are limited accordingly. Put simply, the construct is the attribute being measured by the test, but

this attribute must be defined more precisely than by phrases such as 'English language ability'.

Investigation of the construct is important, as, among other things, it is a way to verify that the interpretations made of the results of the test are valid (AERA, APA, & NCME, 2014). For this to be so, the *de facto* construct of the test must correspond to the definition of the construct interpretations are based upon. In order to verify this, a wide range of aspects concerning the test and its administration and processes must be considered (AERA et al., 2014; Kane, 2006; Saville, 2010). Two concerns involving the construct are of particular importance, however: *construct underrepresentation* and *construct-irrelevant variance* (Messick, 1989). The former concerns parts of the construct which are omitted; the latter, elements which are not intended to be part of the construct but influence test results. Establishing empirically that such concerns are of negligible impact on the test results, and their subsequent interpretation, is the principal aim of investigations into construct validity. Section 1.2 will discuss the context of such studies in order to situate the current study theoretically.

1.1.2 The concerns of the current study

The current study investigates construct validity empirically. Theoretical models of reading were applied to test materials for the Reading component of First Certificate in English (FCE). Information about the materials were collected primarily through machine-based analysis of the texts. This data and the response data from the actual live administration of the test were then modelled statistically to determine which attributes found in the materials had a significant influence on test score. In other words, this study seeks to determine aspects of the *de facto* construct of FCE Reading, at least in respect of the test form examined. The utility of doing this relates to a number of areas of concern in testing generally, including:

- the interpretation of results
- a better understanding of the way in which cognitive processes are affected by contextual features
- the construction of further test forms
- future revisions of test specifications
- automatic generation of test items

The relevance of the current study to these areas will be explained in 1.2, 1.3 and 1.4. The first four offer tangible benefits to test users, as they represent quality improvements. If, in the years to come, it is possible to construct language test items automatically, it will be because of studies like the current one. The most obvious benefit of this is in efficiency savings for test providers but these will, no doubt, be passed on to those paying for the tests. The method adopted is also of note, as it is hoped that it can also be the basis for other studies investigating similar concerns with any language test. In particular, the use of automatic machine generated indices on the test materials (see 1.5) and the use of techniques which allow the application of the method to a single test form. An example of this is the way in which parts of the reading passages are related to test items in order to obtain data for analysis (2.9.2). Furthermore, the psychometric modelling approach used is not common in the language testing field, compared to other methods such as regression. It is, however, straightforward to implement and has many advantages (2.9.3.3.3, 2.9.3.3.4). It is hoped, therefore, that this study will serve to introduce this approach to others for whom it may also be useful.

1.1.3 First Certificate in English (FCE)

As discussed in 1.1.2, the major focus of the current study is on the construct validity of the Reading component of a test of general proficiency in English: FCE. Reasons for this choice will be discussed in 1.6, but this section provides a brief introduction to the test. FCE has components covering each of the four skills (Reading, Writing, Listening and Speaking) and one entitled 'Use of English', which has a lexico-grammatical focus. As FCE targets the B2 level of the Common European Framework of Reference (CEFR) (Council of Europe, 2001) an understanding of what is required of test candidates can be obtained by reviewing the CEFR Can Do statements for B2 and the adjacent ability levels as set out in Table 1. *The First Certificate in English Handbook for Teachers for examinations from December 2008* (University of Cambridge ESOL Examinations, 2007) describes the test as being administered in around 100 countries to a candidature comprising around 200 nationalities, with most candidates aged between 15 and 17. Although the size of the annual candidature is not made public as it is commercially sensitive, since FCE is widespread and long-established, the figure

is significant. For example, the data supplied for this study contained 28,048 candidates from a single test administration.

Table 1 Overall reading comprehension, CEFR levels B1 to C1 (Council of Europe, 2001:69)

Level	Reading descriptor
C1	Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.
B2	Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.
B1	Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.

FCE was first introduced in 1939 and has been revised eight times since then (Hawkey, 2009), not including the latest revision, which was introduced in January 2015. The previous revision was in 2008 and the one before that in 1996.

Perhaps because FCE is such a well-established and internationally recognised exam, it has been used in a number of studies as a kind of baseline to which other tests are compared. For example, Wu (2014) compares FCE to the Taiwanese General English Proficiency Test (GEPT) in as part of a study to establish the alignment of GEPT to the CEFR. Ilc and Stopar (2014) also use FCE as a reference test when examining the link between the Slovenian General Matura Examination (GME) in English and the CEFR.

Among the reasons which make the exam interesting for research are the importance of this exam to so many stakeholders, the attention of the exam owner, Cambridge English Language Assessment (formally University of Cambridge ESOL examinations and before that, UCLES), to maintaining its relevance and focus through research-based revisions (Hawkey, 2009) and FCE's importance as a reference-point for international standards. The December 2005 administration was made available to the researcher, as all its tasks had been retired.

1.2 A paradigm shift in validation studies

1.2.1 The established paradigm

In their seminal paper on construct validity and its research, Cronbach and Meehl (1955) suggested investigating constructs within the test (*studies of internal structure*) and between

tests (*external component of construct validity*), primarily through studies involving the correlations or covariances between test scores. The aim was to establish whether items thought to be testing a specific construct were actually doing so by measuring the way items grouped with each other based on their correlations with other items. If items were expected to be testing the same construct and were indeed found to be closely related to each other, it was considered as evidence of *convergent validity*. By contrast, if those items which were expected to be testing different constructs proved to be relatively unrelated, it was evidence of d*ivergent validity*. The method developed to conduct validity studies in this paradigm was termed *multitrait-multimethod*. In this approach, response data was collected on the performance of candidates on several hypothesised traits, with each trait tested with more than one test method. Among results, convergence and divergence due to traits was distinguished from that due to test methods (Campbell & Fiske, 1959). More recently, Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) have been employed to essentially the same ends: grouping items into a so-called factor structure, which yields evidence of convergent and divergent validity. All these approaches are said to focus on *between-item variance* because the focus is on the differences and similarities of one item to another.

In language testing, the approach to validation suggested by Cronbach and Meehl (1955) has dominated the field for many years. This seems to be in large part because it complemented the nature and aims of frameworks on which language testing researchers based their studies. One such example is the framework Canale and Swain (1980) introduced as *communicative competence* in the early 1980s.

The Canale and Swain (1980) framework was an early example of the application of the communicative approach to language learning. It followed Hymes (1972) and others in its reaction to a narrow view of language, as little more than a system of syntactic rules, promoted by Chomsky (1965). Additionally, it was a reaction to Oller's (1983) view that linguistic competence was essentially unitary, which was based in part on the methodological shortcomings of his own correlation-based studies (Bachman, 2007; Carroll, 1983; Farhady, 1983; Vollmer & Sang, 1983). In comparison to depreciated views of language competence such as Chomsky (1965) and Oller (1983), Canale and Swain (1980) thought it important to delineate what was involved in the broader understanding of linguistic communication. The

result, in the form of Canale and Swain's (1980) model (which was also the basis for later models (Bachman, 1990; Bachman & Palmer, 1996, 2010)), was a hierarchically-structured framework consisting of a set of competences intended to be comprehensive. On the first level of the hierarchy, *Communicative Competence* was divided into: *Grammatical Competence*, *Sociolinguistic Competence* and *Strategic Competence*. In turn, Grammatical Competence comprised the lexicon, morphology, syntax, sentence-level meaning, phonology; Sociolinguistic Competence included the appropriateness and the rules for discourse construction and comprehension and Strategic Competence consisted of strategies required in communication, such as those for repair or to compensate for deficits in other competences.

A nested hierarchy framework such as that of Canale and Swain (1980) lends itself to correlational or factor studies because the nested hierarchy of the theoretical model can be reproduced as a nested hierarchy of factors, thus providing supporting evidence for the theoretical model. Attempts to validate such models empirically were, however, piecemeal (Bachman & Palmer, 1982), and, as Chalhoub-Deville (1997) points out, empirical evidence supporting them has been elusive. Despite this, multitrait-multimethod and factorial studies were used to investigate divisibility of test data, leading, for example, to claims for convergent and divergent validity and, therefore, evidence against Oller's (1983) unitary model. An interest in the divisibility of latent traits, and the use of factorial studies to investigate them continued through the 80s and 90s to the present day (Fouly, Bachman, & Cziko, 1990; Song, 2008; van Steensel, Oostdam, & van Gelderen, 2013). Throughout this period, the between-item factor structure has been understood as a default component in validity studies (Bachman, Davidson, Ryan, & Choi, 1995; Bachman & Eignor, 1997; Sawaki, Stricker, & Oranje, 2009).

1.2.2 An alternative paradigm

Embretson (1983) called for a move away from Cronbach and Meehl's (1955) approach to validation studies to match what she described as a paradigm shift in psychology generally. As she puts it, since the time of Cronbach and Meehl: 'the goal of psychological theorizing has changed from explaining antecedent/consequent relationships to explaining performance from the systems and subsystems of underlying processes' (p. 179). The investigation of the internal structure, became *construct representation*: the determination of the underlying

processes required to complete tasks. In educational and psychological tests, this would mean the cognitive processes elicited by test items and tasks, since these are what the test results could be said to represent.

The tasks involved in educational tests are usually complex and amenable to decomposition into nested sub-tasks, or components, each requiring particular abilities (Embretson & Yang, 2006; Sternberg, 1985). Embretson and Wetzel (1987), for example, hypothesise that two components are involved in the process of responding to items in multiple-choice reading tests: *text representation*, which, in essence, is reading the text, and *response decision*, which is how the candidate determines the response they will select. As most items require both components, but the level of importance of each within each item is expected to vary, investigation concerns *within-item variance*, as opposed to prior interest in between-item variance. In contrast to the correlation-related studies of Cronbach and Meehl (1955) and Campbell and Fiske (1959), construct representation is associated with techniques such as verbal protocol analysis, eye-tracking studies and mathematical or psychometric models (Messick, 1989). Even after Embretson's (1983) paradigm shift, where an interest in within-item variance replaced an interest in between-item variance, studies involving the latter are still common in the literature (Reckase, 2009). Where between-item variance is expected to be much larger, or more significant than within-item variance, this may be an appropriate approach. Such is likely to be the case if the tasks underlying test items are relatively simple and well-defined, so there is little overlap of the underlying processes required for each group of items. Each factor recovered will therefore be distinct. Language tests, which measure the ability to use language, tend not to be among these types of test, however. Language skill performance usually correlates strongly across skills and across tasks within the same skill. As noted in 1.2.1, high correlations between item scores led Oller (1983) to wrongly assume that language ability was a unidimensional trait. van Steensel et al. (2013), for example, attempted to investigate the divisibility of traits within a reading test, where, in their view the three traits tapped by items (retrieving, interpreting and reflecting) are highly distinct. The results of a Confirmatory Factor Analysis (CFA), however, did not support the divisibility of the trait underlying the test. In their conclusions, they suggest that dependency among the hypothesised traits may have contributed unidimensionality within the data.

1.3 Investigating underlying cognitive processes in language testing

Weir (2005) was among the first to present an approach, known as the *socio-cognitive approach*, to validation concerned with the cognitive processes underlying responses to items and tasks. This approach has been adopted by some language test providers, including Cambridge English Language Assessment (Taylor, 2014). A central element of this approach is the decomposition of the processes underlying responses to items into subcomponents. Khalifa and Weir (2009) provide an example of this for the process of reading, and posit the subcomponents listed in Table 2.

Table 2 Subcomponents of the cognitive process of reading (Khalifa & Weir, 2009:43)

	Subcomponent
1	Word recognition
2	Lexical access
3	Parsing
4	Establishing propositional meaning
5	Inferencing
6	Building a mental model
7	Creating a text level structure
8	Creating an organised representation of several texts

This model for reading is more fully discussed in 2.2.1. Distinct models are put forward for each skill, as the underlying cognitive processes are not the same in each case (Geranpayeh & Taylor, 2013; Shaw & Weir, 2007; Taylor, 2011). A corollary of this approach, therefore, is that applying a single model of language ability to all skills, as Canale and Swain (1980) or Bachman and Palmer (2010) sought to do, provides inadequate detail to understand the way in which items differ.

A second key characteristic of the socio-cognitive approach is the need to consider the influence of contextual features on cognitive processing. Table 3 lists relevant categories of contextual features for reading. The features include the *linguistic demands* of the input (reading text), but also other aspects, under the headings *task setting* and *setting: administration*. The influence of context on cognitive processing is illustrated by Khalifa and Weir's (2009) description of the goal setter in their reading model. The goal-setter is an executive function operating during the reading process which involves the determination of

the purpose for reading and thereby the way in which the text is read. Type of reading can involve *careful* or *expeditious* reading at a *local* (small sections of text) or *global* (larger sections of text) level. Compare, for example, reading a personal e-mail from a friend, and browsing a newspaper to determine whether there are any stories of interest. The former is likely to involve, on the whole, careful reading in a sequential manner. The latter, on the other hand, is likely to involve expeditious reading to search for information of interest, perhaps followed by careful reading of a particular story.

Table 3 Context validity (Khalifa & Weir, 2009)

<table>
<tr><th colspan="2">Context validity</th></tr>
<tr><th>Task setting</th><th>Linguistic demands: task input & output</th></tr>
<tr><td>• Response method
• Weighting
• Knowledge of criteria
• Order of items
• Channel of presentation
• Text length
• Time constraints</td><td rowspan="3">• Overall text purpose
• Writer-reader relationship
• Discourse mode
• Functional resources
• Grammatical resources
• Lexical resources
• Nature of information
• Content knowledge</td></tr>
<tr><td>Setting: administration</td></tr>
<tr><td>• Physical conditions
• Uniformity of administration
• Security</td></tr>
</table>

The list presented by Khalifa and Weir (2009), although diverse, is linked by the notion that each element in it may affect the cognitive process of the candidates, and therefore the result. Bachman (1990) presents a similar list of what he refers to as *test method facets*. The influence of each is discussed but, since his approach is not based on the investigation of cognitive processing, the research relating to each facet is not linked in a concerted way. This may be also because Bachman (1990:115) conceptualises them as facets in the same way as would be done in Generalizability Theory. Each is responsible for a certain proportion of variance in the overall test scores. Research would aim to explain overall variance by decomposing it into facets corresponding to contextual features, rather than explaining how these features cause variance by engaging with the cognitive processes of the candidate. The socio-cognitive approach, by contrast, attempts to explain the variance in just this way.

The integral relationship between context and construct means that, in Weir's (2005) approach, the construct is conceptualised as interaction-focussed (Bachman, 2007), or what Chalhoub-Deville (2003) calls ability–in language user–in context. Essential to this conceptualisation is that ability is context-specific. As Chapelle (1998:48) puts it, 'the interactionalist construct definition ascribes observed performance consistency to the combined influence of person characteristics and contexts'. In other words, accurate measurement depends on an understanding of the context within which the candidate is being measured. If the context is changed, the test performance would be expected to alter.

The focus on underlying cognitive processes in language test validation of researchers like Weir (2005), Khalifa and Weir (2009) and Field (2013), is an important new direction for language testing. There are, however, areas which require further development. One of these is the empirical investigation of cognitive processes and the contextual factors which influence them. In recent research, the link between these two elements has been overlooked. Wu (2014) and Ilc and Stopar (2014), for example, use the Weir (2005)/Khalifa and Weir (2009) model to investigate reading tests, by presenting findings on both cognitive processes and contextual features of the texts. However, there is little attempt to determine how the findings on cognitive features are dependent on contextual features: the two sets of findings are treated separately.

In the Khalifa and Weir (2009) reading model (described more fully in 2.2), the relationship between cognitive processing and test performance, or item difficulty, is implied by the hierarchical structure of the subcomponents of cognitive processing listed in Table 2. In other words, subcomponents towards the bottom of the list are considered more difficult. This is mainly because they presuppose the preceding subcomponents and cognitive load is thereby increased. Khalifa and Weir (2009) also describe what they term 'scoring validity', which concerns the measurement properties of item and test scores (e.g. reliability and item statistics). However, they do not seek to describe how the score relates to the cognitive processes. By contrast, in an empirical study, Bax (2013) saw the link between difficulty and cognitive processing as important. He employed eye-tracking and retrospective verbal protocols to identify specific causes of difficulty, such as challenging lexis and syntax, for some items. He then differentiated between the processes used by successful and unsuccessful candidates in his findings. Relating cognitive processes to successful task completion is a key

element of Embretson's (1983) concept of construct representation. Without it, it is impossible to distinguish between contextual effects which might make a significant difference to test results, and those which actually do. The validity of the interpretation of test results clearly depends on this distinction, since contextual effects which have no impact are of little interest.

In addition to the relative paucity of empirical evidence concerning particular contextual effects in language testing, interest tends to focus on only some of the cognitive processes relevant to the testing context. For example, in reading tests, contextual elements connected to the text, such as lexis, syntax and discoursal features are clearly of interest. But the nature of the items could also be expected to have an impact. As mentioned in 1.2.2, in addition to a reading component, Embretson and Wetzel's (1987) model of reading in a test situation includes a response decision component. This component was developed specifically in relation to multiple-choice reading items and comprised stages of falsification and confirmation of options. Contextual effects due to other item types were not investigated, however.

The reasons for the interest of researchers in the target skill alone probably relate to Messick's (1989) conception of construct-irrelevant variance. This appears to be influential in Field's description of a process of validation, where he recommends that a 'model of the target skill as employed by expert users under non-test conditions' is developed and 'the processes which feature in the model [are compared] to the specifications of the test under scrutiny' (Field, 2013:84). The construct is defined in relation to a model based in a non-test situation, cognitive processes and contextual factors which do not match the model are considered construct-irrelevant, and therefore of no interest.

More empirical studies, examining both cognitive and contextual aspects relating to all aspects of completing tests, and relating these findings to relative success on test items and tasks are therefore required. The current study represents one such attempt to fill the gap.

1.4 Other motivations for the decomposition of difficulty

Research into the decomposition of difficulty has not been motivated purely by concerns over construct representation or validity. In fact, in language testing, interest in cognitive processing has rarely been paired with an investigation of features which make a test easier

or more difficult. This in itself is a motivation for the current study. It is, however, worth reviewing other motivations for research into the decomposition of difficulty, as some of them are also relevant to this study.

Research into the decomposition of difficulty can be grouped according to two broad concerns: i) interest in the test and ii) interest in the candidates. In all cases, variables which are thought to influence difficulty were specified and their values determined before some kind of analysis relating these variables to response or difficulty data. Among those concerned with the test, Weir (2013) reviewed indices which could usefully differentiate reading test tasks at defined ability levels. Only linguistic features of the text were included, however – those which map to the components of the Khalifa and Weir (2009) model (see 2.2.1) and, for practical reasons, which are easily available from online sources such as Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2012) and VocabProfile (Cobb, 2013).

Other researchers, for example, Embretson and Wetzel (1987), Rupp, Garcia, and Jamieson (2001), Freedle and Kostin (1993), Gorin (2005) and Gorin and Embretson (2006) note an interest in test development as the context for their research. Embretson and Wetzel (1987) suggest tagging items to be stored in a bank with information about the effect on difficulty of each attribute, something echoed by Weir (2011). The attributes they included in the study were based on a cognitive processing model of responding to items in reading tests. Their model, unlike that proposed by Khalifa and Weir (2009) included construct-irrelevant features. It comprised two sub-processes: text representation (reading) and response decision (selecting a response), which included, for example, a process of determining whether a multiple choice option was falsifiable or confirmable as the key, given the text (see 2.4 for a more detailed explanation of the model).

As with the Embretson and Wetzel (1987) model, Rupp et al. (2001) include text-related and non-text related variables some of which apply to both reading and listening items in their study. Findings were used to suggest changes to the construct definition. Freedle and Kostin (1993), by contrast, state that their main aim is to predict difficulty of the range of defined tasks in TOEFL Reading. The reasons for doing so was due to a number of claims had been made that the multiple-choice tests of reading were unsatisfactory. Most variables were based on the reading passage, and derived from a review of the literature. Some, however,

were related to other contextual features, such as lexical overlap between item stem or option and the text itself. Finally, both Gorin (2005) and Gorin and Embretson (2006) were motivated by the need for information about item constructs for development of automatic generation of items. Their research was based principally on Embretson and Wetzel's (1987) model of reading and involves both text representation and response decision variables. Gorin's (2005) study involved manipulating variables, much as automatic item generation algorithm might be expected to, and then appraising the impact of each manipulation. Gorin and Embretson (2006), on the other hand, decomposed the difficulty of test items using live test data, much as Embretson and Wetzel (1987) had done.

A number of researchers focus on the decomposition of difficulty out of an interest in the candidate. Among these are Carr, Brown, Vavrus, and Evans (1990), Buck, Tatsuoka, and Kostin (1997), Buck and Tatsuoka (1998), Jang (2009) and Sawaki, Kim, and Gentile (2009). In these cases, the aim was to provide diagnostic feedback on test performance. Sawaki, Kim, et al. (2009:190), for example, aimed to 'explore the possibility of developing a detailed score report for low-stakes use'. Although difficulty is decomposed according to the nominated variables, the analytic models used in most such studies are Latent Class Models (LCM), which do not aim to comment on the role each variable has in item difficulty. Rather, the variables are a basis for classification of candidates, so that a detailed skills profile, based on the attributes specified by the model, can be provided to each candidate. Usually, studies involving LCMs seek to provide diagnostic information using a single test. An alternative approach was provided by Carr et al. (1990), who developed a complex model of the cognitive processes believed to be involved in reading, and devised a test battery to measure its components. Feedback instruments were developed based on candidate results on the entire test battery.

1.5 Computer-based recovery of contextual parameters

An important motivation for this study is to inform the practical work of producing tests. For this reason, as with Weir (2013), the use of practical tools which could be applied to test development and construction is considered important. The accessibility of freely-available, computer-based tools for textual analysis facilitates the recovery of indices for studies such as the current one, but also in other aspects of the production of tests. An opportunity exists for the augmentation or replacement of some human judgement with greater consistency

and fewer logistical challenges. Weir (2011), for example, has argued for the use of such indices in vetting texts for test tasks. These tools include those available online, such as Coh-Metrix 3.0 (McNamara et al., 2012) and VocabProfile (Cobb, 2013), which include information from other sources, such as the Latent Semantic Analysis (LSA) website (Laham, 1998) and the British National Corpus (BNC) (BNC Consortium, 2007). It is also possible to go directly to some of these sources (such as the LSA website) to obtain textual analyses. Other tools may be downloaded for academic purposes. One such tool is the Computerized Propositional Idea Density Rater (CPIDR) v5.1 (Brown et al., 2012), which analyses the propositional composition of texts.

For the conduct of the current study, freely-available tools of the kind mentioned offer two major advantages over the analysis of texts by human judges. The first is the time saving involved. Freedle and Kostin (1993), for example, remarked that due to the number of texts in their study, they were unable to attempt a propositional analysis. With a tool such as CPIDR v5.1 (Brown et al., 2012), this may be done instantaneously for each text. It should also be remembered that the more technical the requirements of the analysis, the more specialised training a rater would require and this may be challenging for test developers to provide. The second advantage of using machine analysis of texts is consistency. Unlike human judges, when a machine analyses the same text twice, the same results are guaranteed. These tools, of course, also have disadvantages. They largely produce pre-specified indices, which may not meet requirements precisely. Furthermore, some indices are based on questionable foundations. Weir (2013), for example, expresses doubt about the database underlying the Coh-Metrix index for word concreteness. In his view it is too small, too old and does not incorporate the distinction between morphologically-derived abstract words, and others. Finally, as Weir (2013) and Graesser, McNamara, and Kulikowich (2011) admit, some aspects of textual analysis are not yet possible to do by machine.

Apart from for the purposes of research, other uses have been put forward for the machine-produced indices discussed here. Weir (2011), for example, suggested that they be adopted in the process of item development. Each text which is proposed as the basis for a task can be analysed with tools such as Coh-Metrix 3.0, and compared to parameters previously established by research. The current study is seen as a step towards the operationalisation of

such indices in the test development process: it will help to determine some indices which may be informative for the test which is analysed.

1.6 Study data

1.6.1 Skill to be investigated

Among the four skills, test scores for the so-called productive skills (speaking and writing) are typically generated through a rating process. This usually involves trained experts interpreting the quality of candidate performance in relation to a hierarchy of descriptors presented in the form of a rating scale (ALTE & Council of Europe, 2011). Rating scales target attributes which are relevant to the construct, so the investigation of construct representation is, in some senses, more obvious and straightforward: if the raters perform appropriately, the scales contain the most significant information about construct representation.

In contrast to the productive skills, the receptive skills of Listening and Reading often involve selected response items, or the completion of short answers, both of which are usually mechanically or clerically marked, thus not requiring detailed descriptors of performance. Field (2013:84) claims that listening is 'the most complex of the four skills to test', in large part due to the complexity of the cognitive processing involved. For successful listening, processing must occur within the time frame dictated by the rate of the input, beyond the control of the listener. Reading, on the other hand, is probably the most researched of the four skills, whether for native speakers, or foreign language learners. This extensive research base provides a solid theoretical platform for the investigation of construct representation.

1.6.2 Test to be investigated

The First Certificate in English (FCE) reading component has been selected as the focal test in the current study for a number of reasons. Importantly, as the current study is concerned with the effect of item types on cognitive processes, the pre-2008 revision of the test contained four tasks, each pertaining to their own reading passage(s), and with a different item type. Construct representation studies typically only involve tests with one or two test methods, usually multiple-choice items (see for example, Embretson and Wetzel (1987), Gorin (2005)). In terms of being able to make practical use of some of the findings of the current study, it is important to apply the techniques to a wider range of tasks.

Another reason to focus on FCE is that it is likely to elicit a range of cognitive processes. According to Khalifa and Weir (2009) FCE elicits most of the cognitive processes posited by their model (see 2.2), with the exception of the two highest levels: *creating a text level structure* and *creating an organised representation of several texts*. Wu's (2014) investigation substantiates this, indicating that FCE elicits more higher-level processes (those dealing with the construction of meaning) than the Preliminary English Test (PET), its B1-targeted sister exam. Furthermore, a B2 level test is likely to include a range of candidates in its population (from B1 to C1). As illustrated by Table 1, this will include both those who are limited to reading straightforward factual texts on familiar topics and independent readers of complex texts. Having a range of abilities is important in a study such as the current one, as only data which contains a distinction between candidates on aspects of interest is likely to reveal anything meaningful on these same aspects (Embretson, 1983; Reckase, 2009). The processing demands of FCE go up to 'building a mental model' in the model, which might be expected to challenge some of the weaker readers in the FCE population, and thereby allow such stages to be identified by a study such as the current one. One final reason to select FCE is that it is a popular test, and data sets large enough to conduct sophisticated analysis of the effect of cognitive attributes are available.

1.7 Aims of the study

The aims of this study stem from an interest in the practical validation and test development of FCE. The development of a process which can be applied to the production of other such test forms is, therefore, of equal importance to identifying the relative importance of contextual effects. Such activities are, as Taylor (2014) points out, an important way in which those working within language test providers can contribute to language testing research. The use of predominately machine-derived measures of relevant contextual effects is also important, in large part because the method used to investigate construct representation would not otherwise be practical. The aims of this research are, therefore, as follows:

- To determine elements of the construct representation of the Reading paper of a form of First Certificate in English (FCE) administered in December 2005 (FCE Dec 2005).
- To develop a practical method which can be deployed in the construct investigation of reading tests with varying test methods.

- To trial the use of machine generated indices in the construct investigation of reading tests.

This study does not aim to validate current theory concerning the cognitive process of reading. Such theory is taken as the basis on which investigation of the test is founded. In cases where results do not concur with prevailing theory, it was therefore not be assumed that theory is wrong, as this study is not designed to question the starting point adopted.

1.8 Chapter summary

This chapter introduced a new paradigm for investigating test constructs: construct representation. It involves consideration of the cognitive processes required to respond to items and the contextual factors which influences these processes. Construct representation also aims to relate these processes and contextual factors, which can be seen as attributes of the items, to item difficulty, something not done by all studies investigating test constructs within the new paradigm. Other motivations for investigating item attributes were discussed, as were computer generated indices to measure them and FCE, the test which this study sets out to investigate.

2 Literature review

2.1 Framework for this study

This study involved the linking of item difficulty in a reading test to attributes of the reading passages, the cognitive processes involved and the setting of the test. Since the research was based within a cognitive processing paradigm, an initial stage was to define the *components* and *subcomponents* of cognitive processing which were involved in the test. As discussed in 1.2.2, these components represent stages within a complex task, such as responding to an item in a reading comprehension test; subcomponents were here defined as further subdivisions of the cognitive processes, nested within components. The second stage of this research involved specifying contextual features, or *attributes* nested within each component[1]. These attributes may have been features of text, or any pertaining to the task setting (see 2.7). An alternative approach would have been to examine the literature to determine attributes which predicted difficulty well in other studies, or to carry out an exploratory study that would test all available indices provided by available tools such as Coh-Metrix (McNamara et al., 2012) and ignore overarching theory and componential analysis. This study was, however, interested in determining test variance which can be explained with reference to prevailing theory.

In order to determine the components and subcomponents to be examined, the model of reading proposed by Khalifa and Weir (2009) was adopted as a starting point. This is because a solid theoretical core is necessary for understanding the process of reading, whether it be in a test or non-test situation. There were examples where the lack of a theoretical core meant that attributes for the study were selected on an ad hoc basis. Several researchers (Bachman & Palmer, 2010; Khalifa & Weir, 2009) have discussed the *skills and strategies* approach, which was a conceptualisation of test constructs which dates to the 70s and is derived from pedagogy. It sees reading (and presumably other skills) as capable of being decomposed into *subskills*. For example, *reading for gist* would be one such subskill because it involves different behaviours and purposes from *reading for detail*, say. Conceptualised in

[1] van der Linden (2005:34) defines '*attribute* as a generic term for any property for the design of a test'.

this way, each subskill is *stand-alone*, connections between them are not of particular interest and the influence of contextual factors is not considered. Subskills are a convenient typology into which behaviours or items may be classified, new ones can be devised if the existing set proves to be insufficient.

Even though the skills and strategies approach dates to the 70s, modern day research replicates its faults. Jang (2009), for example, described a process of determining item attributes which involved raters examining a number of the verbal protocols of trial candidates, among other sources. The result was a list of nine 'reading skills':

1. Context-Dependent Vocabulary Skill
2. Context-Independent Vocabulary Skill
3. Syntactic and Semantic Linking Skill
4. Textually Explicit Information Skill
5. Textually Implicit Information Skill
6. Inferencing Skill
7. Negation Skill
8. Summarizing Skill
9. Mapping Contrasting Ideas into Mental Framework (MCF) Skill

As Alderson (2010) points out, Jang's list is different from that of Sawaki, Kim, et al. (2009), even though they analysed the same test. Some of the skills relate to contextual aspects of the text (e.g. negation skill), some more to the requirements of responding to items (e.g. summarizing skill).

The ad hoc nature of these categories limits what can be said about the test being researched, as it is hard to compare them to the results of similar research of other tests (or even the same test). Furthermore, it is doubtful that a typology so test- and research-specific could be of much diagnostic assistance for learners. A cognitive processing approach, based on a theory of reading, however, would be more likely to yield findings comparable with other research because it would enable the comparison of specific cognitive processes and contextual features. It would also better equip the researchers to diagnose areas of improvement for learners which are not test-specific, as the processes in question would be

found outside the testing context. The next section will describe the theoretical starting point of the current research.

2.2 A cognitive processing model of reading

2.2.1 The Khalifa and Weir (2009) model of reading

Khalifa and Weir (2009) propose a cognitive processing model of reading, illustrated in Figure 1. According to Zwaan and Radvansky (1998:162), 'language is now seen as a set of processing instructions on how to construct a mental representation of the described situation'. The central spine of the model in Figure 1 contains the stages by which these instructions are decoded and implemented. The three initial boxes at the bottom represent what some researchers call lower order processing skills, in contrast to the higher order skills (Weir, Hughes, & Porter, 1990), which are placed in the upper portion of the spine. Lower level processing produces the building blocks (e.g. the meaning of words, their syntactic relations) for constructing the overall sense of the text or parts of the text, which is the concern of higher level processes. A dependency exists between those stages further up the spine, and those below them. Because the effect is cumulative, each progressive stage implies a more difficult cumulative challenge to any reader. The ability of the reader being adequate, the extent of progress up the spine of the model in Figure 1 by any reader is dependent on the demands of the reading being done. For example, the creation of an intertextual representation is only required where more than one text is being read.

2.2.1.1 Lower level processes

Word recognition, according to Khalifa and Weir (2009:47) involves 'matching the form of a word in a written text with a mental representation of the orthographic forms of the language.' A word may be recognised as a whole (the *lexical route*), or through breaking the word into *sub-lexical* elements and determining the relationship between graphemes and phonemes. This latter route is relatively difficult for learners of English, due to the complex interrelationships between sounds and graphical forms in the language (Khalifa & Weir, 2009). Once a word form has been matched, the reader attempts to attach a semantic form to it. This stage is termed *lexical access* in the model and relies heavily on the extent of the reader's mental lexicon. Finally, syntactic parsing involves both supra- and sub-word grammar, and seeks to determine the syntactic relationships between words within a sentence or clause.

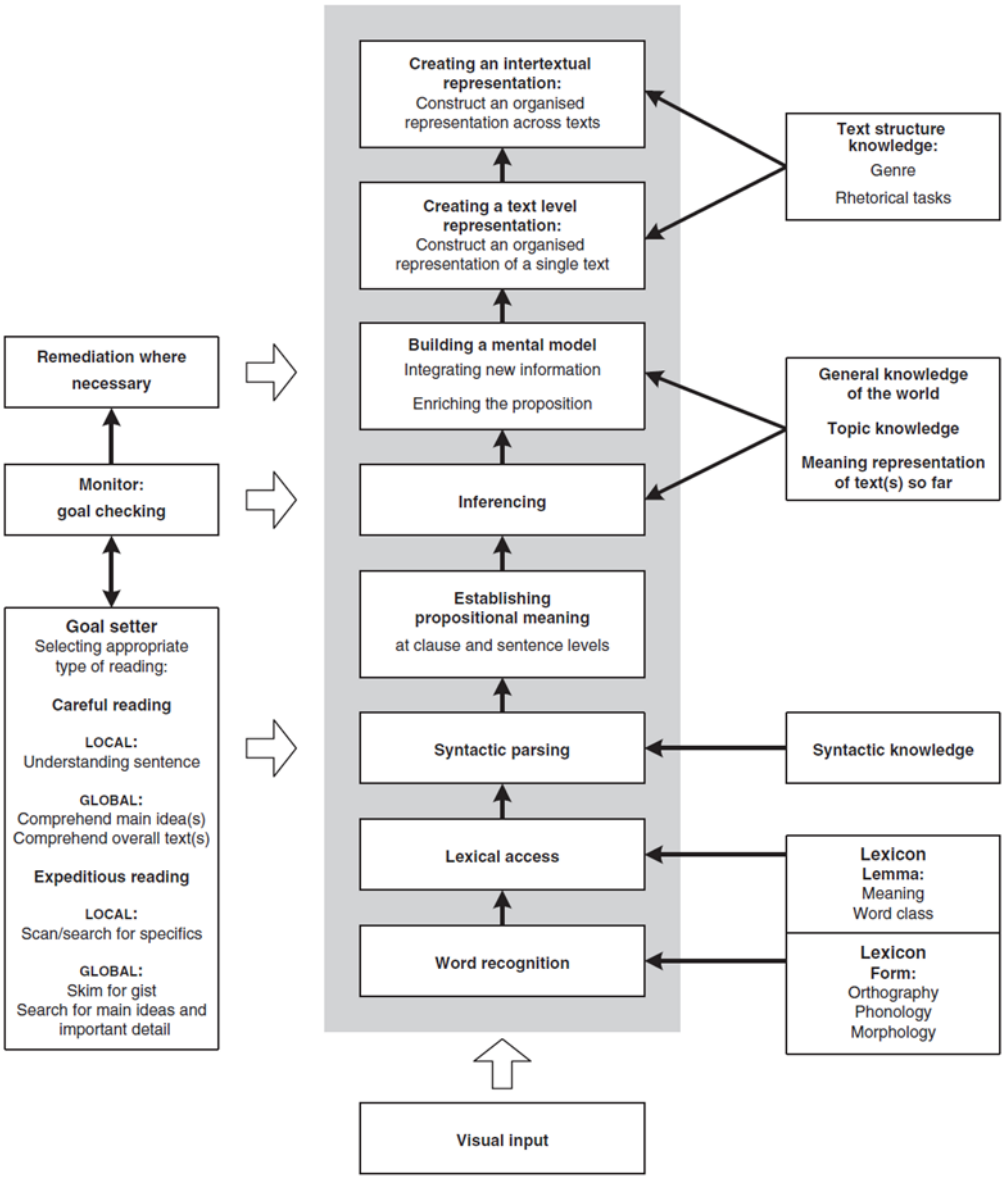


Figure 1 Model of reading (Khalifa & Weir, 2009)

2.2.1.2 Higher level processes

The first stage of meaning construction (fourth box from the bottom, central spine, Figure 1) is establishing propositional meaning. Khalifa and Weir (2009:50) define this as 'a literal interpretation of what is on the page'. The activity at this stage is probably best illuminated by considering Kintsch and van Dijk's (1978) notion of a textbase, which is a mental representation of the propositions interpreted up to that point by the reader. The propositions referred to are described as 'idea units' by Zwaan and Singer (2003) and consist of one predicate (for example, an action: 'jump' in 'Frank jumps') and one or more arguments

(the agent of the action: 'Frank'). The textbase usually remains in the short-term memory for seconds, until the information it contains has been integrated with the working mental representation of the situation being described in the text, or *situational model* (Zwaan & Singer, 2003). Comprehension at this stage, of course, relies on lower order processes such as word identification and parsing. For fluent reading, these lower order processes must be possible with the minimum of conscious effort (Grabe, 2009).

The next stage in the Khalifa and Weir (2009) model is termed *inferencing*. It is characterised by the introduction of prior knowledge in order to develop links between elements of the text which are not explicit, and usually known as *bridging inferences*. Zwaan and Singer (2003:100) provide the following example,

> The lightening struck. The hut collapsed.

The causal relationship between the two propositions is not stated but may be needed if the remainder of the text (not given) is to be coherent. Inferencing is also said to include the determination of the meaning of unknown words from their context, as well as anaphor resolution, where words which refer to other words (such as pronouns) are linked. In all these cases, the object is to render a coherent understanding of the text.

After Inferencing, building a mental model[2] comprises integrating the information processed so far into a mental understanding of the situation described (Zwaan & Radvansky, 1998). Creating a text level representation involves the construction of discoursal representation of the whole text, where different propositions (micro- and macro-) stand in hierarchical relation to each other. This representation dovetails with Khalifa and Weir's (2009) distinction between global reading, which concentrates propositions near the top of the hierarchy, and are spread throughout the text, and local reading, whereby the reader aims to comprehend propositions at all levels of the hierarchy within a limited range of text. The final stage involves the construction of a similar representation across more than one text.

To the left of the central core of the model in Figure 1, the executive functions are represented. According to Khalifa and Weir (2009:44) 'decisions taken on the purpose for the

[2] *Mental model* and *situational model* are treated as interchangeable in the current study. The latter term will be adopted to avoid confusion.

reading activity will determine the relative importance of some of the processes in the central core of the model', or, in other words, the type of reading. Careful reading is the processing of text more or less sequentially and with the monitor 'at a high level of attention' (Urquhart & Weir, 1998:107). This may be over a limited local range, or globally. Such attention levels are not required for expeditious reading, including skimming for gist, scanning for very specific information, or search reading for information on topics of interest. Careful global reading requires all steps of the central core, up to at least *creating a text level representation*. Expeditious reading may require no higher order processing at all. According to Urquhart and Weir (1998), scanning to find a name will involve limited lexical access and syntactic parsing. Skimming, on the other hand, will involve the extraction of a textbase macrostructure. For Urquhart and Weir (1998), this does not necessarily imply the creation of a situational model or a text level representation but, conceivably, this may be done, if required. Urquhart and Weir (1998:108) characterise search reading as a 'search for information on prespecified macropropositions'. They further comment that, for skimming or search reading, location of relevant text will trigger careful reading.

The final column of the model (to the right) includes the mental resources likely to be required at each stage. Resources corresponding to lower level processing are linguistic in nature. Those corresponding to higher levels also demand world knowledge for Inferencing, and more specialised types of knowledge, such as of text genre.

2.2.2 The Khalifa and Weir model and FCE

The model of reading proposed by Khalifa and Weir (2009) is relatively new but two studies have been conducted employing this framework to investigate FCE. Wu (2014) compared FCE to the Taiwanese General English Proficiency Test (GEPT) in order to examine the alignment of the latter to the CEFR. In order to do so, experts were asked to categorise items according to the subcomponents found in the Khalifa and Weir (2009) model which each test was thought to elicit. She found that higher-order processes were more common in FCE than its B1 sister test, Preliminary English Test (PET). However, higher level processes were less frequent, generally speaking, than lower level processes (Figure 2).

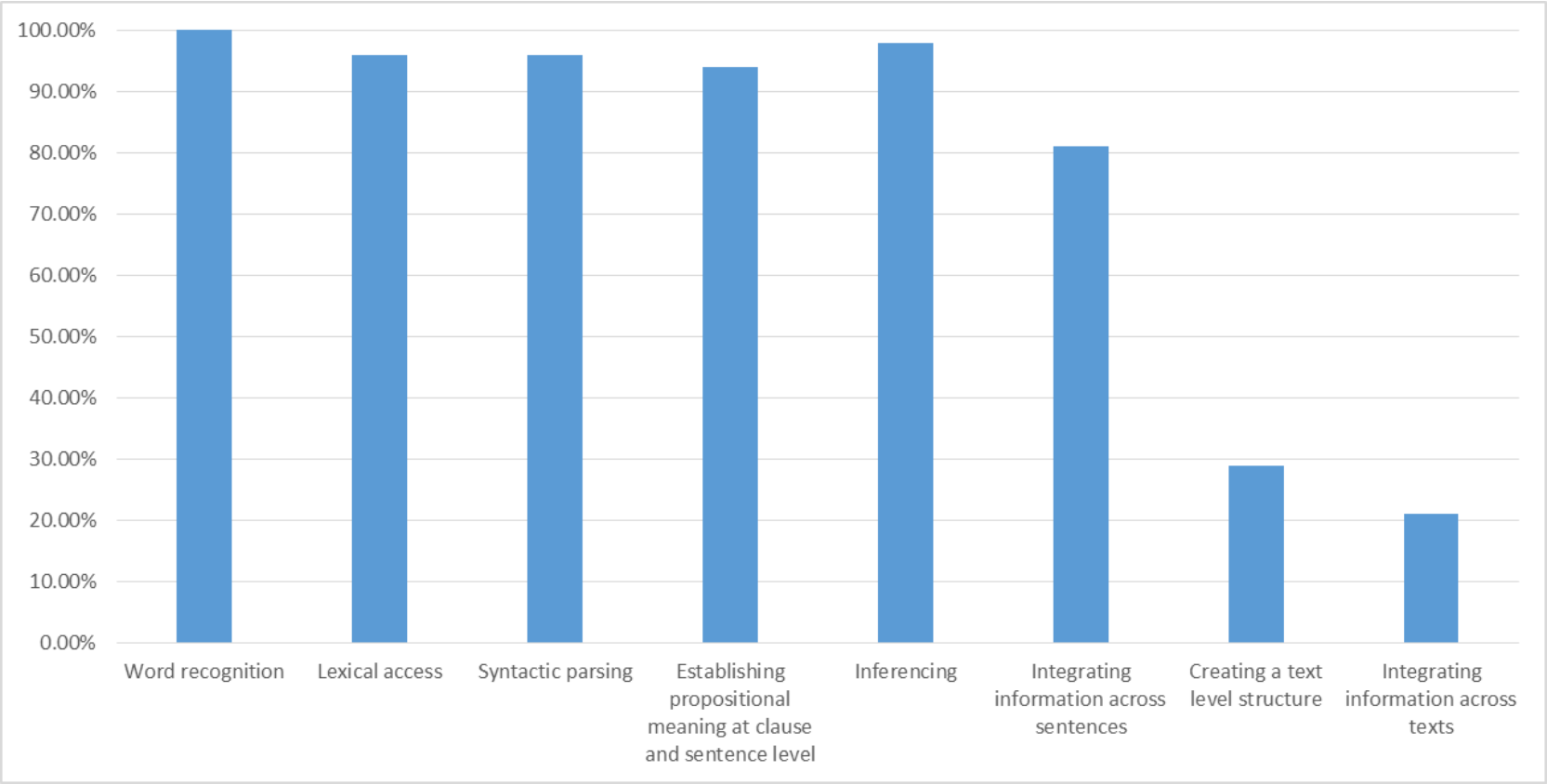


Figure 2 Results from expert judgement on cognitive processes (FCE) (Wu, 2014:112)

Contextual factors were also examined by Wu (2014), although this was done in isolation from investigation into cognitive processes. Machine analysis of the reading texts was done using tools such as Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2005) and the results compared between the tests examined.

In a similar fashion to Wu (2014), Ilc and Stopar (2014) examined the cognitive processes elicited by FCE, as well as contextual features, in order to determine the alignment of two other tests, the General Matura Examination (GME) A and B, to the CEFR. Their findings concerning FCE correspond closely to those of Wu (2014), with a similar pattern of demand on lower and higher level processes. Ilc and Stopar (2014), however, do not report the result for FCE directly, only a summary of their findings.

As the studies by Wu (2014) and Ilc and Stopar (2014) are so similar, methodologically, they share the same advantages and disadvantages. The use of expert judgement to categorise items according to cognitive process implies comprehensiveness when considering how elements of the Khalifa and Weir (2009) model are represented among the items. This is because judges considered a small and finite set of categories equating to the entire range of cognitive processes implied by the theoretical model. As a result, it was possible for both Wu (2014) and Ilc and Stopar (2014) to present the relative frequency for each subcomponent for the whole test (Figure 2). In studies, such as the current one, where sources of item difficulty are identified empirically, the same comprehensiveness is not possible. The categories used by Wu (2014) and Ilc and Stopar (2014) are not directly observable, so must be inferred from what is. Undetected sources of difficulty are inevitable, and these will be manifest as measurement error (Kane, 2011).

There are three principal disadvantages in the approach followed by Wu (2014) and Ilc and Stopar (2014). First, the cognitive processes involved in responding to

items are recorded, but their impact on test performance (see 1.2.2 and 1.3) is merely assumed to correspond to the underlying theoretical model, rather than investigated empirically. Such an assumption is too strong when empirical data is available to test it, as was the case with both studies. Second, the impact of contextual features on test performance was not explored (1.3). Without knowing how contextual features relate to cognitive processes, only a partial picture of how the test functions is recovered. These two limitations are partly related to the choice of method used to determine the cognitive processes elicited: expert judgement. Judges could have been asked further questions about the relation of cognitive processes to difficulty, or the role of contextual features but would have stretched the limits of their expertise. Instead, an empirical approach which interrogates live test data would have been more suitable. This is an important reason why the current study follows and empirical, rather than judgemental approach. This approach will be elucidated in the remainder of the current chapter. The third drawback found in Wu (2014) and Ilc and Stopar (2014) is that they are insufficiently critical of the model proposed by Khalifa and Weir (2009). Specific limitations of the model are discussed next (2.2.3).

2.2.3 Difficulties with the Khalifa and Weir (2009) model

The most significant difficulty with the Khalifa and Weir (2009) model is that it does not contain mechanisms to explain how the executive functions interact with other cognitive processes and contextual features, or the way in which contextual features interact with cognitive processes. These features are all described and examples are given of how they work, but this is done on an ad hoc basis and the model itself is not referenced. For example, despite explanations of types of reading in Khalifa and Weir (2009) and Urquhart and Weir (1998), the cognitive process which triggers one type of reading over another is described in the model. Similarly, although descriptions make it clear that contextual factors are important influences over cognitive processes, the model does not contain a mechanism which would allow a researcher to predict how a feature such as text complexity might affect reading. As a consequence, for studies such as the current one, where

empirical evidence is sought, a way must be found to supplement the model if it is to be used.

For the purposes of the current study, a further problem with the Khalifa and Weir (2009:43) model was that it is designed to account for the process an expert native speaker would adopt in a non-test situation. Its intended use is as a template with which to compare the processes elicited by a test (Field, 2013). For this reason, some processes associated with the test are not included in the model and an understanding of the way they work must be sought elsewhere. For example, the process of selecting a response to a particular item type is not a feature of the model.

A final issue with the Khalifa and Weir (2009) model concerns inferencing. The model foregrounds inferencing, making it a stage in the cognitive process. In the Khalifa and Weir (2009) model, each stage requires those preceding it (i.e. those at a lower position in Figure 1). Seeing inferencing as a stage in its own right is problematic because it is not always required by later stages. Depending on the reading purpose, the text may not be read carefully and linearly and, whether inferences were required would depend on the nature of the text which is actually read and the information which must be extracted. In a test situation, particularly, candidates may only read short segments of text (Rupp, Ferne, & Choi, 2006). Parts of the text where a bridging inference is required may be avoided, but the construction of a coherent textbase might, however, still be required for successful performance. A more productive way of seeing inferencing is as a means, not always required, to the end of establishing a coherent textbase, and also, sometimes a requirement in building a situational model. The stage in question should therefore be renamed 'establishing a coherent textbase', as this is always required before building a situational model.

2.3 Activating the goal setter: Rouet's (2012) TRACE model

In order to investigate the way in which the goal setter and contextual features interact with the cognitive process, it was necessary to augment the Khalifa and

Weir (2009) model. As discussed in 2.2.3, their model does not contain a mechanism to explain how this is done. Furthermore, processes which may be expected during test taking, such as choosing from among alternative responses to an item, are not included in the Khalifa and Weir (2009) and had to be added for the purposes of the current study.

Rouet's (2012) model was adopted to provide the aspects missing from the Khalifa and Weir (2009) model. The two models are both based on the belief that the influence of contextual factors and characteristics of the individual, such as prior knowledge, are important influences over the process (Khalifa & Weir, 2009; Rouet, 2012), which makes integrating them easier. Although the TRACE model does not attempt to describe cognitive components of reading in the detailed way of Khalifa and Weir (2009), it contains scope for such detail to be added. Rouet's (2012) model has also been validated empirically in a number of studies. When, for example, it was used to investigate computer-based reading, it successfully explained search patterns and text structure recall based on question specificity and prior knowledge (Rouet, 2003). Rouet, Vidal-Abarca, Erboul, and Millogo (2001) also found that search patterns were influenced by contextual aspects – specifically, the cognitive load induced by the items.

Rouet (2012:105) proposed the Task-based Relevance Assessment and Content Extraction (TRACE) model to be applicable to 'any situation where the reader's purpose is to gather information that fits a pre-existing need'. In test taking, the pre-existing need is to respond correctly to test items. As the TRACE model has such a general purpose, included in the description which follows is consideration of how it may be adapted to the purpose of reading in a test situation. The model, illustrated in Figure 3, is divided into four main parts, *info-based processes*, *information resources*, *memory-based processes* and *memory resources*. The resources support their respective processes; processes relating to information use external documents as the main input/output, whereas memory-based processes interface with the reader's internal resources stored in the memory.

The process begins with 'examining the initial set of constraints that motivate the subject's activity' (Rouet, 2012:106), which could be reading the rubric and the stem and first option of a multiple-choice item (1). The central spine of the Khalifa and Weir (2009) model, may be considered a good representation of the cognitive processes at this point. This results in (2), the construction of a *task model*, or a 'representation of the actions to be performed in order to complete the task' (Rouet, 2012:106), which is stored among the reader's memory resources. The task model helps to define the goal setter for the task and activate the monitor which will remain active until the phase involving this sub-task is complete, or the task model is updated. This process determines whether (3) *external information* will be sought, which is very likely in the case of a language test. However, the reader may decide that prior knowledge is sufficient at this stage to (7) *update the internal response model* and, if in (8), the *response* is considered *complete*, the candidate will *output the task product* (9), or, in a test, mark the chosen response as required by the test according to the *response model* stored in the memory.

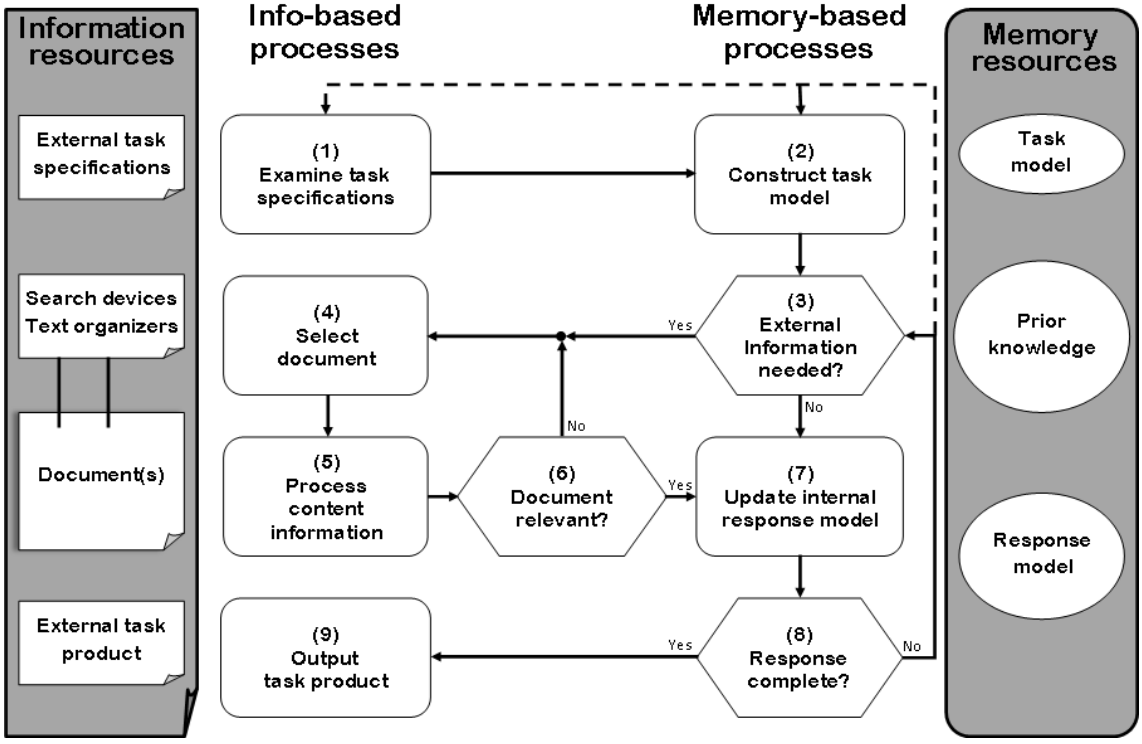


Figure 3 Task-based Relevance Assessment and Content Extraction (TRACE) model (Rouet, 2012:105)

After the task model is constructed, the reader may decide to (3) seek external information. If so, the model requires that (4) a *document* be *selected*. Once promising text is located, (5) the *processing* of *content information* is required. At this point again, models of reading such as that proposed by Khalifa and Weir (2009), are relevant. Once the selected text is processed, a decision must be made as to whether *document* is *relevant* (6). The search-process loop can be repeated as often as is necessary until it is determined that sufficient information has been obtained to move to (7) and then on to (8). If it is decided that, at (8), the response is incomplete, the reader can go back to decision (3), or further back, to the beginning of the process.

One element of the TRACE model needs to be adapted slightly to allow it to fit a language test. However, this adaptation has more to do with the way in which is described than of significant difference in the process. The point of the model to which it applies is (4). In many language test reading tasks, only one text is provided so this stage appears unnecessary. However, it is described as involving search techniques based on the task model. For this reason, for the purposes of the current study, 'document search' was substituted with the 'search for relevant text', which may be a segment of the text provided.

Gaps in the Khalifa and Weir (2009) model are filled by Rouet's (2012) model in the following ways. First, the goal setter in the Khalifa and Weir (2009) model can be understood as the task model of the TRACE model. In an item-based task, its construction explicitly involves characteristics of the items, as such, the rubric or stem of an item is considered to be the input. For other types of test task, the task model is still applicable. Tasks involving summarising reading texts, for example, are given with instructions, such as to write 'a report on the central ideas of the source text for classmates who had not had a chance to read the source text themselves' (Yu, 2008:530). These instructions provide the basis for the task model when reading. For item-based reading tasks, the language used in the stems and the conceptual demands of the task described by the rubric may be more or

less difficult. This would affect the difficulty of constructing an appropriate task model, which, in turn would have ramifications for completion of the remainder of the task. Second, in the Rouet (2012) model, stage (4) was interpreted as searching for text relevant to an item. As such, the reasons for a reader to employ different types of reading become clear. The reader must search for text which is relevant to the task model for which they have previously formulated a model. The likely type of expeditious reading at this stage may be predicted from the input to the task model. So, for example, if the task model requires the location of a name, scanning may be adopted, whereas, if some form of specific information is needed, search reading is more likely to be used.

2.4 Adding construct-irrelevant contextual factors: Embretson and Wetzel's (1987) General Information-Processing Model for Multiple-Choice Paragraph Comprehension Items

As discussed in 2.3, the TRACE model presents a framework for the entire task of reading and allows the relationship between the executive processes, other cognitive processes and contextual factors of reading to be modelled. By combining this model with the detailed description of reading provided by Khalifa and Weir (2009), a better understanding of cognitive processes during a reading test may be obtained. Other parts of the process, however, specifically those involving the selection of the response are not described in sufficient detail. A final model, that produced by Embretson and Wetzel (1987) provides a good starting point for modelling this part of the process.

The model used by Embretson and Wetzel (1987) to investigate the processes within passage-based multiple-choice reading comprehension tests is illustrated in Figure 4. It consists of two components: *text representation* and *response decision*. The former comprises *encoding* and *coherence processes*, the first of which is largely based on Kintsch and van Dijk (1978) and can be understood as synonymous with Khalifa and Weir's (2009) first three stages of lower order processing plus the establishment of propositional meaning. *Coherence processes* are the formation of propositions into a coherent textbase.

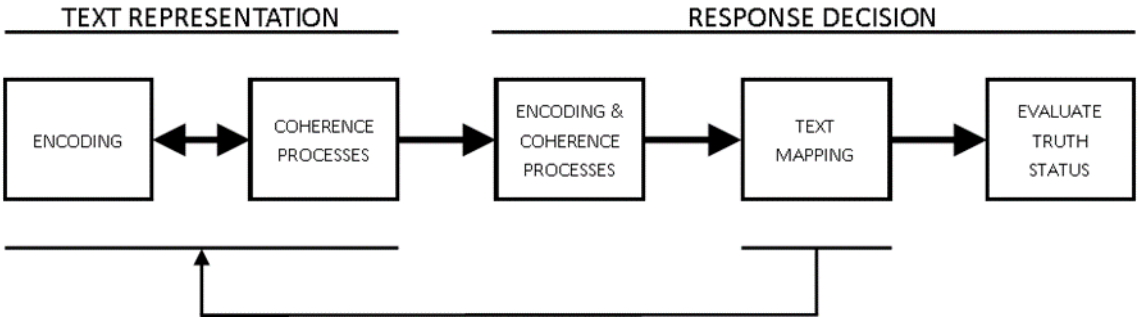


Figure 4 General Information-Processing Model for Multiple-Choice Paragraph Comprehension Items (Embretson & Wetzel, 1987)

The second component of the Embretson and Wetzel (1987) model, *response decision*, is specific to the process of selecting the response to test items. *Encoding and coherence processes* here are the same as those under *text representation* but applied to the text of the item stem and options. *Text mapping* is the process of locating appropriate text for each option (similar to Rouet's (2012) step 4), so includes a recursive loop to *text representation*. The final component, *evaluate truth status*, involves determining which alternative to select for the response (like Rouet's (2012) step 8).

Embretson and Wetzel (1987) specify the final component, evaluating the truth status, as having two stages: *falsification* and *confirmation* (Figure 5). This involves the candidate first attempting to reduce the number of options under consideration by rejecting some as distractors according to the evidence available in the text. Following this, the candidate attempts to choose between those options remaining by comparing the available supporting evidence for each.

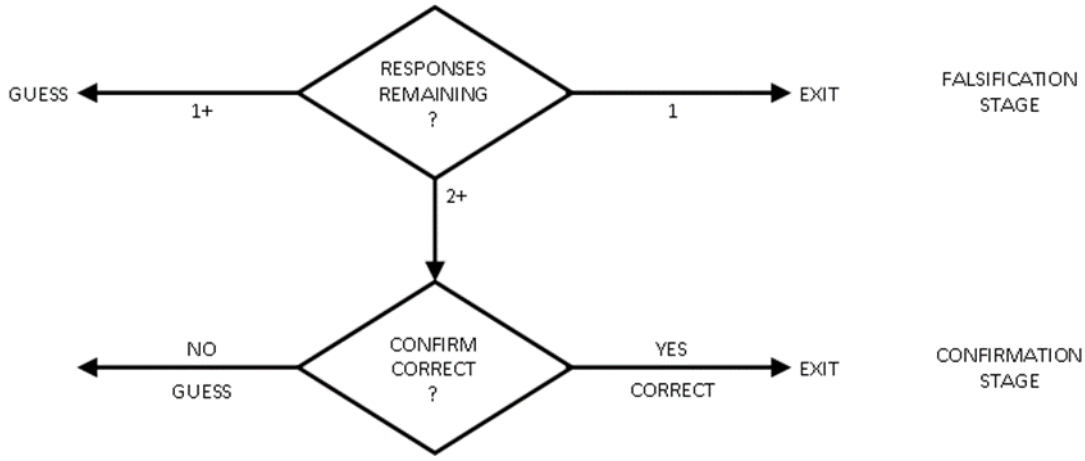


Figure 5 An information-processing model for evaluating the response alternatives (Embretson & Wetzel, 1987)

Compared to the Khalifa and Weir (2009) model, Embretson and Wetzel's (1987) offers a narrower view of the cognitive processes concerning reading, as it presupposes only careful, local reading. Furthermore, the cognitive process is not as extensive as that modelled by Rouet (2012), omitting consideration of the construction of a task model. The response decision component of the Embretson and Wetzel (1987) provides something unavailable in the other models.

Embretson and Wetzel's (1987) response decision model has been the subject of some criticism from Rupp et al. (2006). They argue that 'a logical process of elimination of incorrect distractors' (p. 446) is problematic because it precludes other ways of selecting responses. A candidate's remark in a retrospective interview from their own research on multiple-choice items supported the case for a more flexible model. The candidate said that she embarked on a process of falsification of options when she decided she had limited understanding of an item but went straight to the confirmation stage in other cases. In other words, for items perceived as easy, only the confirmation stage was used. For items perceived as difficult, the falsification stage was used. This would suggest a greater influence of all options for difficult items, and a reduced influence for all options but the key in an easy item.

The criticism presented by Rupp et al. (2006) does not represent an alternative response model to that of Embretson and Wetzel (1987) but rather an example of how the model may not apply under certain conditions. The particular strategy adopted by the candidate in Rupp et al.'s (2006) study may only be applicable where some items are written in such a way that only one option is plausible. Furthermore, their study was not conducted using a live test for which the candidates had been preparing independently. It may be that the candidate felt willing to use such a strategy where the stakes for her were quite low and an unfamiliarity with the test had meant she did not possess alternative approaches. Nevertheless, Rupp et al. (2006) do show that the model proposed by Embretson and Wetzel (1987) may be too simplistic for wholesale adoption. In a more general sense, however, two ideas implicit in the model are of value. First, that contextual aspects of the key and the distractors are important for the selection of the response, and second that falsification and confirmation can play a role in the selection of a response. For the current study, therefore, contextual features of the key and distractors were examined independently to determine, whether they influence item difficulty. The comparative impact of key and distractors were of interest, as they could potentially shed light on whether the falsification stage is more influential than the confirmation stage.

2.5 Formulation of a theoretical composite model

For the purposes of the current study, the key elements of the models discussed above were combined into a composite model. The model is, of course, an idealisation of the process of reading in a test, and excludes, for example, recursive loops between components which are permitted in Rouet's (2012) model, as these would make the model far more complex and, therefore, more difficult to implement. Only those parts of the model which can be implemented in a relatively straightforward way will be discussed further. Table 4 lists the components of a theoretical composite model.

Table 4 Components for a theoretical composite reading model

Theoretical component	Theoretical basis	Description
Task model	Rouet's (2012) *task model* together with Khalifa and Weir's (2009) model of reading	The candidate formulates their approach to the item, considering, in particular, the text of the item stem and/or option
Search	Rouet's (2012) *select document* together with Khalifa and Weir's (2009) types of reading	The candidate searches for relevant text to read more carefully
Meaning construction	Khalifa and Weir's (2009) model of reading together with Rouet's (2012) *process content information*	The candidate processes the information found to construct meaning
Response decision	Embretson and Wetzel's (1987) *response decision* together with Rouet's (2012) *response model*	The candidate determines which option to select as a response

2.6 Operalisation of the composite model

The operalisation of the model is yet another stage of idealisation. Exactly what is meant by each component was defined in relation to the subject of study. Once components were defined, subcomponents and attributes which nest within them were defined.

2.6.1 Task model/OP

The task model must be operationalised with reference to a specific text. This is also the case with the meaning construction component. A difficulty arises, however, due to the some of the test methods found in FCE Reading, 2005. This is because the input for the task model could be either the text of the option, or the main text. This is particularly true for Part 1, where the reading text relating to each item is clearly demarcated and a list of headings is provided as options (see

Appendix 1: test papers) – some candidates may prefer to start with the text, others with the options. And since this is a problem for the task model, there is a corresponding issue for the search and meaning construction components. The candidate will search the text they did not use to form the task model and what they find will be the input for the meaning construction component. In reality, Part 1 is likely to elicit a highly complex pattern of reading, where components like the task model are constantly refined using the main text as well as the options. However, since this was not the focus of the current research, and because such a pattern would be impractical to model, significant generalisation would have to be made about how the task model is formed for Part 1 and also for Part 3, which has a similar task format.

In the case of the multiple-choice items which relate to a single text in Part 2, formulation of a task model based on the item stems seems reasonable. This is because, unlike the demarcated text of Part 1, reading the main text alone will not provide candidates with any specific information about the task they are required to undertake. Part 4 is rather similar to Part 2, where reading the main texts alone do not provide the reader with an understanding of the task. Using the item stems as input for the task model would also seem reasonable in the case of this task, therefore.

Because of the uncertainty in Parts 1 and 3 associated with whether the item stems or main text is used to form the task model, it is not possible to nominate attributes which unambiguously relate to the first three components. For this reason, a more pragmatic approach was adopted, whereby for all Parts, attributes which relate to the item stem and options text was said to define the task model component, and attributes relating to the main text, the search and meaning construction components. The changes to the theoretical composite model are summarised in Table 5. For the task model component, item difficulty is hypothesised to influence item difficulty through the ease with which the associated text may be read. For this reason, attributes related to the stem and

option text which were hypothesised to influence reading difficulty were incorporated into the overall model. These are discussed in 2.7.1. It is important to note that the text for the OP component would normally be expected to influence item difficulty less than the text for the READ component (discussed in 2.6.3). This is because, generally speaking, the OP text helps the reader to form a task model, which guides their reading of the main reading passage. If the OP text is too difficult, the relevant part of the reading passage may never be accessed, and the item would test something quite different than intended. To some extent, therefore, difficulty in the OP text may be considered construct-irrelevant variance.

Table 5 Components for an operationalised composite model

Operationalised component	Associated with	Theoretical component	Theoretical basis	Description
OP (options and item stems)	Item stem and options text	Task model	Rouet's (2012) *task model* together with Khalifa and Weir's (2009) model of reading	The candidate formulates their approach to the item, considering, in particular, the text of the item stem and/or option
SEARCH	Item stem and options text and the main reading text	Search	Rouet's (2012) *select document* together with Khalifa and Weir's (2009) types of reading	The candidate searches for relevant text to read more carefully
READ	Selected text from main reading text	Meaning construction	Khalifa and Weir's (2009) model of reading together with Rouet's (2012) *process content information*	The candidate processes the information found to construct meaning
RD (response decision)	Selected item stem and options text and associated selections from main reading text	Response decision	Embretson and Wetzel's (1987) *response decision* together with Rouet's (2012) *response model*	The candidate determines which option to select as a response

2.6.2 SEARCH

This component links the text of the options with the main reading text. In terms of item difficulty, its influence is understood to be the facility with which relevant

text can be found after a task model has been formed. This may include the strength of the match between the option text and the main text, and other attributes related to the nature of the test method. These attributes are discussed in 2.7.2.

2.6.3 Meaning construction/READ

As mentioned in Table 5, this component is associated with the main reading text. As with task model, aspects which affect the difficulty of reading were considered as attributes here, and will be discussed further 2.7.1.

2.6.4 Response decision/RD

The response decision concerns the process of selecting an item response. Attributes must involve, therefore, the linguistic features of the text involved and their configuration within the key and distractors of the item. For example, this can include the strength of the link between the key and its associated text, or that between the distractors and their associated text. Such attributes will be discussed 2.7.3.

2.7 Specifying subcomponents and attributes for components

As the current study was an investigation of construct representation, attributes included are hypothesised to affect the difficulty of an item. Determination of attributes is not the search for universals, however. According to Bejar (2010:215),

> the reading process as part of testing [is] a very specialized form of reading, but the reading required by different assessments could well be different, and, therefore, the same models of difficulty perhaps should not be expected to generalize completely.

Diverse findings are not uncommon in studies involving the investigation of test-related attributes. Gorin and Embretson (2006) attempt to fit two attribute-based models (Embretson & Wetzel, 1987; Sheehan & Ginther, 2000), developed for specific tests (the Armed Services Vocational Aptitude Battery (ASVAB) and the Test of English as a Foreign Language (TOEFL), respectively), to data from the Graduate Record Examination (GRE). The results of the analysis explained only

moderate levels of the variance, however: 29% for the first model and 25% for the second.

Gorin (2005), in her study involving the manipulation of attributes, also found a disappointingly thin relationship between attributes posited by other studies to predict item difficulty, and the empirical difficulty of GRE items. She cited Jackson's (2005) finding that the models do not necessarily fit data from populations for which they were not developed. An explanation for the difficulty of generalising such model may be found by considering the dimensional structure of data, within which attributes are nested. Dimensions are not, as is sometime assumed, a property of the test but of the data, which is a record of the interaction between the candidates and the items (Reckase, 1994). As also noted in 1.6.2, dimensions, and by extension, attributes, therefore, will only be recoverable from data if the candidates have shown a range of performance over the dimension (Reckase, 2009), or the attribute. This logic can also be applied to the other side of the interaction which produces the data: items. If the items do not require the involvement of a particular attribute for successful completion, candidates will not be afforded the opportunity to show a range of performance over the attribute and it will again not be recoverable from the data. For example, *text to correct option lexical overlap* (word spots) was determined to be a significant predictor of difficulty on TOEFL reading by Freedle and Kostin (1993) but such overlap may be totally absent from other tests, resulting in this attribute being uninformative.

Next, the sources of difficulty, which can be considered as contextual attributes, and their operationalisation, will be considered. They will be presented as nested within their subcomponents, which are in turn nested within their components.

2.7.1 OP and READ

According to the composite model adopted for the current research, reading is important when forming the task model and when constructing meaning from the reading passage. For this reason, the stages derived from the Khalifa and Weir (2009) model were treated as subcomponents in each case (see 2.2.1). The first

four subcomponents were shared by both OP and READ, therefore. Since OP is operationalised as reading the stem and option text only, and sometimes not in the form of complete sentences, subcomponents above that of establishing propositional meaning were not considered relevant. For this reason, the final three subcomponents exclusively relate to the READ component.

2.7.1.1 Word recognition

The first stage in the Khalifa and Weir (2009) model is word recognition. As noted above, words can be decoded either through recognition of the word as a whole, or by identifying elements which make up the word. . According to Weir (2013), it has been shown that the number of syllables affects processing time and therefore of difficulty. For this reason, to the extent that word recognition is conducted via decomposition into syllables, the number of syllables will affect difficulty. Readers may also recognise a word as a whole without needing to decompose it. Nevertheless, the mean number of syllables per word was selected as an indicator of difficulty because it was expected to represent difficulty at this stage.

2.7.1.2 Lexical access

Lexical access comprises assigning meaning and other characteristics to words based on the contents of the mental lexicon. Unknown words, or words not known well to the reader are likely to cause difficulty at this stage. A reader might be expected to have greater knowledge of more frequent words and little or no knowledge of less frequent words. For this reason, measures of word frequency are expected to be indicative of difficulty.

A large corpus, like the British National Corpus (BNC) (BNC Consortium, 2007), which contains more than 100,000,000 words, is likely to provide the most useful measures of difficulty and was employed by Weir (2013) through an online tool called VocabProfile (Cobb, 2013). Words are grouped in strata according to their frequency, such that the most common 1,000 words form the first group, followed by progressively less frequent strata. Group membership of words in the text of interest can be counted, and higher totals for strata of less common words can be expected to indicate more difficult items. The Academic Word List (AWL)

(Coxhead, 1998) is also suggested by Weir (2013) and is also available online (Cobb, 2013). It can be used in a similar way to the BNC strata. The list comprises 570 headwords and 2,570 words in total which occur more frequently in academic texts and are not found in the first 2,000 most frequent words in English as listed by West (1953). As academic words are thought to be more difficult than the first few thousand most common words in English, VocabProfile divides input text into those in the first 1,000 most common words, those in the second most common 1,000 and those on the AWL.

Frequency measures from one further corpus are also recommended by Weir (2013): the Centre for Lexical Information (CELEX) database of word frequencies, which consists of 17.9 million words (McNamara, Graesser, McCarthy, & Cai, 2014). McNamara et al. (2012) provide CELEX frequencies in three ways: an index representing the frequency of content words, the logarithm of the frequency of content words and the logarithm of the frequency of all words. The logarithm is used because, according to Graesser et al. (2011) and McNamara et al. (2014), this provides a linear relation to reading times. In other words, reading times increase exponentially as frequency decreases.

Other relevant lexis-related attributes are polysemy and hypernymy. The former measures the number of distinct meanings a word has. For example, *chair* may be a kind of seat, or the leader in a meeting. The more senses a word has, the more difficult it may be to correctly match information contained in the mental lexicon and the less clear the meaning of a proposition may be (McNamara et al., 2014). However, frequent words typically have more senses, so polysemy does not necessarily indicate difficulty (McNamara et al., 2014; University of Memphis, 2012).

Hypernymy can be thought of as a measure of word specificity. This is because hypernyms have a more general meaning in comparison with their subordinate words (compare *vehicle* with *car*, *bus*, *lorry*, and compare *car* with *hatchback*, *estate*, *limousine*). In some cases, more specific words are expected to be known

by learners and knowledge of their hypernyms is a mark of more sophisticated leaners, as in the first example. However, the relationship between hypernymy and proficiency is non-linear: *car* is likely to be better understood by most learners than *hatchback*. Coh-Metrix (McNamara et al., 2012) provides indices for both polysemy and hypernymy based on content words[3] matched in WordNet (Fellbaum, 1998), a database containing information about the relationships between more than 170,000 words (McNamara et al., 2014). A higher polysemy score indicates more polysymous words, whereas a higher hypernymy score indicates more specific words.

Lexical density is here defined as the ratio of content words to function words. A higher proportion of content words means that more lexical resources must be devoted to retrieving the meaning and other information about the words in the sentence (lexical access). In contrast, according to Weir (2013), function words require less processing, as they may be anticipated and skipped; furthermore, their frequency and typically shorter length results in easier recognition (Weir, 2013). Coh-Metrix (McNamara et al., 2012) includes incidence indices for nouns, verbs, adjectives and adverbs. To construct the index of lexical density for this thesis, these were summed and divided by the total number of words, also an index available through Coh-metrix.

The concreteness or abstractness of the words involved is also a consideration. According to Weir (2013:525), abstract words are harder 'to process because they are not as imageable as concrete words'. Presumably, this refers to ease of lexical access. It seems likely that word concreteness may also be a productive indicator of difficulties in constructing the situational model. If concepts are more abstract, the ease with which the situation described by the text can be understood would be expected to be reduced. Coh-Metrix (McNamara et al., 2012) provides an index of word concreteness, based on human ratings of 4,293 words (McNamara et al.,

[3] Content words are defined verbs, nouns, adverbs and adjectives. Non-content words are known as function words and include pronouns, prepositions, determiners and conjunctions.

2014; University of Memphis, 2012). Weir (2013) argues that the studies upon which the index is based (Coltheart, 1981; Paivio, Yuille, & Madigan, 1968) are, to some extent, questionable, as they are old and methodologically and theoretically deficient. For Weir (2013), the size of the initial word list was small (925 words), as well the failure to deal with the distinction between abstract terms which are formed morphologically, such as 'happiness', and those which are not, such as 'truth' (p.525). Two further criticisms may be levelled at these indices. First, they may not include a rating for every word in the text being analysed but no explanation is given by the developers of Coh-Metrix about how missing data is dealt with. Second, the relative concreteness of words which are far apart is likely to find agreement among raters. However, differential ratings of words which are perceived as near are not only likely to elicit less agreement but also to beg the question of whether real differences can only be understood with a specific context. The index was included in the current study, despite its shortcomings, for two reasons: limited alternatives existed and its empirical performance could be used to determine whether it is of benefit to the study.

2.7.1.3 Syntactic parsing

Parsing consists of classifying words as parts of speech and grouping words into meaningful blocks, such as noun or verb phrases. Syntactic complexity is thought to be the major source of difficulty at this stage because classifying and grouping words depends on recovering their role in the sentence. With more complex sentences, roles are harder to determine.

The complexity of individual syntactic units is expected to have a direct effect on the difficulty of syntactic parsing. According to Weir (2013), the occurrence of more modifiers (e.g. adjectives) per noun phrase increases the cognitive load during parsing. Similarly, a larger number of words before the main verb increases cognitive load (University of Memphis, 2012) while the reader is locating the main verb (Weir, 2013). The location of the main verb of the main clause is seen as key to parsing because other constituents of the sentence may be identified in relation to it. A larger number of words to its left are thought to increase difficulty because

they must be processed before the main verb is identified. In addition, the length of the noun phrase will increase cognitive load because more words must be parsed Weir (2013). Coh-Metrix (McNamara et al., 2014; McNamara et al., 2012) offers indices both for the number of words before the main verb (which they refer to as *left-embeddedness*) and the number of modifiers per noun phrase,.

2.7.1.4 Establishing propositional meaning

Establishing propositional meaning involves reconstructing relationships between elements of the text. Propositions consist of a predicate, which relates elements, and at least one argument. Kintsch (1998:38) provides the following example:

- Predicate: give
- Arguments:
    - Agent: Mary
    - Object: book
    - Goal: Fred

Which is the propositionalised form of the sentence:

Mary gave Fred the book

One of the elements which can make comprehending the relationships between elements of a proposition difficult is negation. Negation has long been thought to contribute to difficulty in reading (e.g. Freedle & Kostin, 1993). According to Weir (2013:510), the difficulty is due to the need to reverse 'a positive concept in order to construct a negative one'. As such, it is semantic in nature, and does not only encompass grammatical or morphological negation (such as the inclusion of the word *not*, or the morpheme *un-*, as in *unhelpful*), but also words which indicate a negative concept, such as lack of something, e.g. *paucity* or *deficit*. Coh-Metrix (McNamara et al., 2012) includes an index for 'negation expressions' (University of Memphis, 2012) but this does not include semantically-based negation. For this reason, an additional, more holistic index, was constructed to count all instances of negation.

What Freedle and Kostin (1993) call *fronted structures* are also likely to cause difficulty for readers. Fronted structures are identified by non-standard word order at the beginning of a sentence. They may include, for example, cleft sentences (*it is here that it all began*), sentences with marked topics (*after a long, hard life, he died*) or combinations of the two. When included in a text, such structures may aid cohesion, helping readers and benefitting comprehension. However, to weaker readers, they may be difficult to interpret because the reader cannot rely on standard syntactic patterns. The reader will require additional cognitive resources to understand the relationships between elements of sentences that include these structures. This implies difficulty in parsing, establishing propositional meaning and establishing a coherent textbase (the second example above contains three propositions which must be related to each other).

Freedle and Kostin (1993) felt that fronted structures were more likely to hamper the efforts of the readers in their study, and this, indeed, was among their findings. Hawkins and Buttery (2012) name some types of pseudo-cleft sentences among the criterial features which define productive capacity at B1 and B2. As a result, it seems reasonable to assume that, at B2 level, cleft and probably all fronted structures have still not been fully internalised by learners. For these reasons, in the current study, an attribute to include fronted structures was included with the initial expectation that it would contribute to the difficulty of establishing propositional meaning and/or syntactic parsing.

Passive voice is a further feature expected to affect the cognitive load of readers (University of Memphis, 2012). Propositional meaning, and possibly parsing and establishing a coherent textbase, is expected to be more difficult when the agent is not explicitly identified within the sentence. Coh-Metrix (McNamara et al., 2012) provides an incidence index for agentless passive voice.

It is worth noting here that fronted and passive structures, in another classification scheme focussing more broadly on syntactic complexity, might have been grouped

together with left-embeddedness and number of modifiers per noun phrase. Shiotsu & Weir (2007), for example, found the broader concept of syntactic complexity to have a significant effect in their data. The categorisation of these attributes in the current research was made on the basis of the cognitive processing stages each seems more likely to have a stronger impact.

2.7.1.5 Establishing a coherent textbase (the attributes in this section were only applied to the READ component)

The textbase is organised into a hierarchical system, whereby propositions differ in their importance to the aims of the text as a whole (Kintsch, 1998). The textbase may need to be supplemented at this stage to establish coherence. That is, the reader may need to infer propositions not explicitly stated, based on the propositions which are stated and their prior knowledge. Khalifa and Weir (2009) refer to this stage as inferencing but this is considered an unsatisfactory characterisation for the reasons given in 2.2.3.

The number of propositions to be contended with in a text is a consideration, as well as their density per word. More propositions will require more resources for processing in an absolute sense, as will more propositions within a specific segment of text, which will increase the relative difficulty of that segment. Brown, Snodgrass, Kemper, Herman, and Covington (2008) advocate the propositionalisation of texts based on part of speech tagging. This approach is based on the association of propositions with verbs, adjectives, prepositions and conjunctions (Covington, 2012). Automatic propositionalisation offers a huge advantage in time and effort over the work of trained human raters, as it requires an extremely detailed decomposition of texts. Over 80 texts, Brown et al. (2008) report a correlation of 0.9693 between human raters and an earlier version of the software they designed for the purpose, CPDIR (Brown et al., 2012). The current version, 5.1, was used for the analysis in this thesis.

A number of textual devices which facilitate the organisation of the textbase into a hierarchical system. Connectives, which provide an explicit link between parts of the text, are particularly important in this respect. University of Memphis (2012)

define connectives as comprising five types following Halliday and Hasan (1990): causal (*so*), logical (*and*), adversative/contrastive (*although*), temporal (*until*) and additive (*moreover*). A higher number of explicitly stated connectives is expected to enhance coherence and make building a situational model easier.

In addition to connectives to link parts of the text, repetition and co-reference (where two words refer to the same entity, or one refers to the other) also increase cohesion. According to Weir (2013:519), it 'reinforces current themes' but also expedites lexical access, presumably because the meaning of the word does not need to be retrieved from long term memory after the first time it is read. Despite this, it is the presence of links between parts of the text which are considered the most important influence of reading difficulty. Coh-Metrix (McNamara et al., 2012) provides indices for the incidence of connectives and for overlap between sentences. The latter index is termed *stem overlap*, and counts nouns which share lemmas with any words of an adjacent sentence. McNamara et al. (2014:50) provide the example of 'swimmer' and 'swimming' (verb), found in consecutive sentences. Although each is a different part of speech, they possess a common root.

Lexical diversity within the text of interest can also be considered a source of difficulty at the textbase stage. The addition of any new information is likely to add to the processing time needed (McNamara et al., 2014). However, according to McNamara et al. (2014), it would additionally be expected to influence difficulty in lexical access. This is because a greater range of lexis will require a larger mental lexicon and the processing time to access new words, whereas the counter-point of less diversity would probably be paired with more redundancy and, therefore, greater cohesion. A number of indices are available to measure lexical diversity but the ratio of *types* (instances of specific words) to *tokens* (all words), or the *type-token ratio* (Weir, 2013), is perhaps the most intuitive to interpret, and will be adopted as the measure of lexical diversity throughout the remainder of this thesis. This index, however, can be problematic. Alderson (2000) argues that the

index is relatively crude and Castello (2008), for example, notes that the index varies considerably by text length. The input texts for each component in the current study are generally short they do not greatly vary in length for their respective component (see Table 10 for specific figures). Furthermore, any variation that there may be due to text length is not likely to obviate its value for the study. It is the absolute difficulty of reading each input text, regardless of length, which is of interest, as this is what is expected to affect the difficulty of each respective test item.

A final attribute of interest for this subcomponent is the number of sentences contained in the relevant text. This is thought to be indicative of whether local or global reading is required. Khalifa and Weir (2009:59) define careful local reading as 'processing at the decoding level until the basic meaning of a proposition is established… [but not] integrating each new piece of local information into a larger meaning representation'. If the text relevant to an item contains more than one sentence, linking them into a coherent textbase would be entailed. Global reading taxes mental resources to a greater extent than local reading because, in order to assemble information into a coherent whole, each part must be stored in the short term memory. Furthermore, connections between sentences must be realised, usually with the help of features such as references or lexical cohesion. The number of sentences is something of an approximation for this source if difficulty, however, as more than one proposition can occur within a single sentences, and cohesive devices can be used to link elements within sentences as well as between them. The number of sentences therefore constitutes a narrow view of the distinction between local and global reading.

2.7.1.6 Building a situational model (the attributes in this section were only applied to the READ component)

A situational model is the distillation of the situation described by the text, so that it is remembered largely free of the surface elements of the textbase. Five situational dimensions are put forward by Zwaan and Radvansky (1998). During the process of reading, they argue, at least five situational dimensions are salient

to readers as they process a text (although this will obviously depend on the characteristics of the text and reading purpose): space, causation, motivation, protagonists/objects and time. As the reader processes text, they may find that they are, at some points, unable to establish continuity within one of these dimensions by their interpretation of the explicit meaning in the text. For an expert native speaking reader, this may be because the information is not stated in the text. However, for other readers, including foreign language learners, they may simply have difficulty decoding the text they have read. In either case, the reader will attempt to re-establish coherence by calling on prior knowledge to make a bridging inference (Zwaan & Singer, 2003).

Coh-Metrix (McNamara et al., 2012) provides indices relating to three of the five situational dimensions posited by Zwaan and Radvansky (1998). These are causal cohesion, or the ratio of causal particles to causal verbs, intentional cohesion, or the ratio of intentional particles to intentional actions or events and temporal cohesion, which is based on the repetition of tense and aspect. Particles are words or phrases which signal the relationship between parts of the text (*because*, *in order to*, *before*). A higher ratio of particles to verbs is expected to aid continuity for situational dimensions (Graesser et al., 2011). The index for the temporal dimension assumes that continuity in tense and aspect equate to continuity in the temporal dimension, with higher scores indicating fewer shifts (Graesser et al., 2011).

2.7.2 SEARCH

Two attributes expected to affect the difficulty of searching are:

- whether the order of information in the text parallels the order of the items
- whether the relevant text for an item is demarcated in some way

Where items are placed in correspondence with the order of the relevant information in the reading passage, according to Khalifa and Weir (2009), search time is reduced as the candidate progresses through the task. However, Khalifa

and Weir (2009) see this kind of item-text correspondence as mainly of concern in relation to the candidates' ability to form a situational model of the entire text. If the items are in a jumbled order, the way candidates approach the text, and consequently the situational model they form, may be quite different from that formed in a non-test situation. In some contexts, though, such as reading for academic purposes, where expeditious reading precedes careful reading (see Weir, Hawkey, Green, & Devi, 2012), they concede that such a pattern may be more appropriate. Clearly, where relevant text is demarcated, search time is further reduced or non-existent, for example, by specific reference to part of the passage, or where a gap in the text indicates missing text. In both cases, once the relevant text has been identified, these indices may be derived by the application of a rule[4].

A third attribute may also be specified for SEARCH. The act of searching is likely to be facilitated by the closeness of the match between the surface features which are used to generate the task model (e.g. stem and option of a multiple-choice item) and the relevant text which is the target of the search. Freedle and Kostin (1993) found that content words in the item text which were the same as, or lexically related to, content words in the part of the passage containing information relevant to the key made the item easier and this may be due to the facilitating of a search for the relevant text. Latent Semantic Analysis (LSA) provides a means of comparing two elements of text on the basis of their semantic content (Landauer, McNamara, Dennis, & Kintsch, 2011). A computer algorithm is trained on a larger number of texts and develops a network of relationships between words based on the context of their occurrence in the training texts. The algorithm relies entirely on semantic relationships and does not consider grammatical similarity at all. LSA has been found to work well, for example, in completing matching tasks within TOEFL (Landauer & Dumais, 1997). A web-based tool is available which provides an index which described the latent semantic match

[4] Instances are counted – a clerical task based on a definition of what counts as a single instance.

between two texts (Laham, 1998). It is possible to create the index by matching the individual words of the text to each other (term to term), or by matching the overall meaning of the texts (document to document). The former was used in the current study, as this would appear to replicate the process of searching, where only surface features would be matched and careful reading of the text would not be expected.

2.7.3 RD - response decision

The idea put forward by Embretson and Wetzel (1987), that there are two sub-components to the decision making process when selecting a response, provides a way to understand how the interaction between the characteristics of distractors and the key contribute to item difficulty: the attractiveness of the key is balanced against that of distractors. Rupp et al.'s (2006) finding that some candidates may simply select what they believe to be the correct response, without considering the others extensively, simply indicates that some candidates may determine that the balance is overwhelmingly in favour of one of the options. The attractiveness of options is usually termed 'plausibility'.

Plausibility is considered to be overwhelmingly semantically-based, and there have been several attempts to model it. Freedle and Kostin (1993) and Kirsch and Mosenthal (1990), for example, both consider it to be related to the overlap between the text of the distractor and the reading passage. Semantic overlap is a key feature in both cases but Freedle and Kostin (1993) go further, including attributes such as the relative position of words and the closeness of the form of the words in stem and text. Embretson and Wetzel's (1987) approach involves the notion of falsifiability and confirmability. If a distractor is judged to be falsifiable, or the correct response deemed confirmable according to the text of the reading passage, difficulty is expected to decrease. Both Rupp et al. (2001) and Gao and Rogers (2011) use the number of options judged to be plausible and find a significant influence in each case. This approach is, however, somewhat circular when plausibility is seen as an empirically-verifiable outcome of the test taking process. A distractor can be found to be empirically plausible when more of the

total number of candidates select it, and therefore do not select the key. When fewer candidates select the key, the item is more difficult according to quantitative measures of difficulty (i.e. in Classical Test Theory the facility value will be lower, or in Rasch measurement, the difficulty value will be higher). Furthermore, defining plausibility by the number of plausible options would not explain an issue that must be at the heart of any study into construct representation: the question of *why* some options are more plausible than others.

Judging the key to be confirmable is relatively straight forward when multiple-choice items are being considered. However, with other item types, such as multiple matching, the response to the item involves more than the consideration of the option and the text. Options and the subject text must be compared for the best match. Clear confirmation from the text is highly unlikely because the item would be considered poor if only one option was plausible. The match between the key and the relevant text, as well as that between the distractors and their relevant text was therefore taken as a measure of relative plausibility. A good match between key and text was expected to lead to easier items, whereas a good match between distractors and text was expected to make items more difficult. LSA was used in this study, in addition to a term to term match, document to document was also specified. Term to term was expected to correspond to the process of referring back to the surface features of the text, whereas, a document to document match was expected to correspond more closely to reference to the situational model of what is described.

During the response process, it is expected that the options and the reading passage will be revisited several times. This may mean re-reading text directly, accessing the situational model formed by the reader or both. In light of this, the configuration of the reading passage is thought to be important. Gorin and Embretson (2006) refer to Kintsch's (1998) theory that information in close proximity in the text is also more closely linked in the mind. Retrieving information so linked is easier; information which is dispersed is more difficult to retrieve. It

may also be the case that, in revisiting relevant parts of the physical representation of the text, an element of expeditious reading is still required, although probably greatly reduced from what was required to initially identify the segment in the search component. For reading in a test, the theory predicts that the closer all the information concerning the plausible options of one item is, the easier the item should be. This was operationalised as an index by counting the total number of words from the first instance of the information concerning plausible options of the item, to the last, according to its placement in the text. The number of words which were judged relevant to key or options was divided by this figure.

Finally, the ease with which a response decision is made is expected to be influenced by familiarity with the relevant text. The more times relevant text has been encountered in working through previous items, the lower the allocation of resources required in subsequent processing of the same text. This may be due in part to the extent to which lexical access of the same words is expected to be facilitated by repetition. Once the relevant text was determined for each option, the number of reoccurrences of each sentence for subsequent options was recorded and an index constructed.

2.8 Considerations in operationalising the composite reading model

The model defined and operationalised in 2.5, 2.6 and 2.7 is summarised in Table 5 and in Table 6, Table 7, Table 8 and Table 9. The model had a hierarchically nested structure, such that information obtained about any element (attribute or subcomponent) nested in another (subcomponent or component) also comments on the secondary element. There was, however, no expectation that the findings of this research would be comprehensive. In other words, as with any statistical analysis, the measurement will involve error, most notably in the form of imperfect specification of some attributes. As a result, whatever can be concluded about a component or subcomponent should be seen as a partial representation - a complete representation could only come with a complete and perfectly specified set of attributes. For this reason, the current research was not be able to

conclude the exact composition of each subcomponent and component, only whether some of the hypothesised attributes did indeed contribute to the difficulty of obtaining scores on items and therefore provide an indication of the influence of each subcomponent and component.

Table 6 Operalisation of the composite model - OP

Subcomponent											
Word recognition	Lexical access	Syntactic parsing	Establishing propositional meaning	Establishing a coherent textbase	Building a situational model	Item order	Demarcation	LSA match	Option match	Dispersal	Practice
Syllables per word	Lexical frequency (BNC, AWL, CELEX) word knowledge (polysemy, hypernymy, concreteness) lexical density	Modifiers per noun phrase left embeddedness of main verb	Negation holistic negation fronted structures passive voice								

Table 7 Operalisation of the composite model - SEARCH

Subcomponent											
Word recognition	Lexical access	Syntactic parsing	Establishing propositional meaning	Establishing a coherent textbase	Building a situational model	Item order	Demarcation	LSA match	Option match	Dispersal	Practice
						Correspondence between the order of items and the relevant text	The level of demarcation of relevant text in the reading passage	The match between the relevant text for an option and the text of the option			

Table 8 Operalisation of the composite model - READ

Subcomponent											
Word recognition	Lexical access	Syntactic parsing	Establishing propositional meaning	Establishing a coherent textbase	Building a situational model	Item order	Demarcation	LSA match	Option match	Dispersal	Practice
Syllables per word	Lexical frequency (BNC, AWL, CELEX) word knowledge (polysemy, hypernymy, concreteness) lexical density	Modifiers per noun phrase left embeddedness of main verb	Negation holistic negation fronted structures passive voice	Number of propositions proposition density number of connectives type-token ratio	Causal cohesion intentional cohesion temporal cohesion						

Table 9 Operalisation of the composite model - RD

Subcomponent											
Word recognition	Lexical access	Syntactic parsing	Establishing propositional meaning	Establishing a coherent textbase	Building a situational model	Item order	Demarcation	LSA match	Option match	Dispersal	Practice
									Key -relevant text match (term to term) Distractor-relevant text match (term to term) Key -relevant text match (document to document) Distractor-relevant text match (document to document)	Dispersal of relevant text in number of words	The reuse of relevant text for later options

2.8.1 Complexity of attribute and component network

As is evident from the discussion in 2.7 concerning attributes, both the components and the attributes are highly interrelated. For example, the task model component is not easily separable from the read component because, in Parts 1 and 3 in particular, the source text for each could be derived from either the main reading text or the text of the options. In the case of the attributes, the type-token ratio is hypothesised to affect lexical access and establishing a coherent textbase. Furthermore, when language use is considered, it is clear that processing at any stage in models like that of Khalifa and Weir (2009) are dependent on all the preceding stages. This was noteworthy because it had implications for the analytical method, which must allow easy separation of elements of the model (discussed in 2.9.3.3.1) and for the interpretability of the results (see the assumption of absence of colinearity 2.9.3.3.4, 3.8.2.2, 3.8.3.1.1).

2.9 Analytical methodology

2.9.1 Data

2.9.1.1 Test materials and response data

The data for this study comprised responses from 28,048 candidates to 35 FCE reading items, administered in December 2005, with a time given of one and a quarter hours. The test consisted of four tasks, referred to as 'Parts', each with a text or texts and several items and outlined in Table 10; the test papers are available in Appendix 1: test papers.

Table 10 Outline of FCE Reading, December 2005

Part	Number of texts	Total text length (words)[5]	Text description	Number of items	Item type	Instructional rubric
1	1	767	An article about fitness and exercise	7	Multiple matching (8 option)	Choose from the list **A-I** the sentence which best summarises each part (**1-7**) of the article. There is one extra sentence which you do not need to use.
2	1	697	A newspaper article about a musical family	8	Multiple-choice (4 option)	For questions **8-15**, choose the answer (**A, B, C** or **D**) which you think fits best according to the text.
3	1	533	An article about a bird called the kingfisher	7	Multiple matching (8 option)	Eight sentences have been removed from the article. Choose from the sentences **A-I** the one which fits each gap (**16-22**). There is one extra sentence which you do not need to use.
4	4	568	A magazine article in which various people talk about their jobs	13	Multiple matching (4 option)	For questions **23-35**, choose from the people (**A-D**). The people may be chosen more than once.

[5] For Parts 1 and 3, this figure includes text which was described as removed from the text and must be reinserted by the candidates.

The FCE Reading test was revised in 2008. Khalifa and Weir (2009), summarise the type of reading and cognitive processes expected in the revised version of FCE Reading. This revised version contains tasks that are apparently equivalent to Parts 2 to 4 in the earlier version that is the focus of this study: Table 10. Their summary of the kinds of reading process that they consider to be required in responding to the tasks is provided in Table 11. In the 2005 test, Part 1 is similar to Part 3, in that external text must be matched to sections of the reading text. Table 10 and Table 11 illustrate the diversity in the tasks contained in the paper (both before and after the 2008 revision). However, in Part 1, the external texts are headings for each of the paragraphs in the main text. As stated in 1.6.2, an important reason for selecting FCE was its diversity of tasks: the inclusion of contrasting test methods increases the likelihood of finding task-related impacts on performance.

Table 11 Attributes of sample FCE Reading paper (Khalifa & Weir, 2009:64-5,72)

Equivalent 2005 Part	Type of reading	Cognitive processing
2	Careful reading global: tests the candidate's ability to identify main points in a text, involving inferencing in a number of questions and one (item 8) relating to a large section of the text.	Usually requires integration of new information sometimes across large sections of the text (see item 8). Many of the answers require the reader to form inter-propositional connections (e.g. items 2, 3 & 5).
3	Careful reading global: tests the candidate's ability to identify main points in a text, involving inferencing in a number of questions and one (item 8) relating to a large section of the text.	Requires integration of new information. No need to create a text level structure because sentences rather than paragraphs are being inserted. In order to complete the task successfully, candidates need to use the clues provided by, for example, discourse markers, understand how examples are introduced and changes of direction signalled. This often needs to be combined with inferencing, e.g. in item 10, where candidates need to realise that putting up tents in muddy fields is not seen as glamorous (see also items 9, 12, 13).
4	Expeditious local and occasional global reading: tests candidates' ability to locate specific information in a text or a group of texts.	Mostly only requires understanding sentence level propositions to answer the questions once the information has been located (see however item 18). May involve inferencing in those items which test understanding of attitudes or opinions (see item 21).

In addition to response data derived from the administration of the Reading test in December 2005, as discussed above, items were coded according to a range of

attributes. The result was a second data set, referred to as an *incidence matrix*. The principal method of analysis related the variables which represented item attributes in the incidence matrix to the difficulty of the items, according to the response data set.

2.9.2 Sampling of text

In 2.7, a number of attributes related to the test are described. In order to derive variables, test materials have to be processed in some way. In other words, text must somehow be analysed to yield a variable which records the number of relevant propositions per item.

Direct judgement of characteristics by experts, has often been employed to identify attributes in similar studies (e.g. Buck et al., 1997; Jang, 2009; Shiotsu, 2010; Wu, 2014) but was rejected for all but two indices in the current study. The two indices where it was used were those of holistic negation and fronted structures, as these attributes were thought potentially important but no machine-derived indices were available. There were two principal reasons to reject direct judgement of indices in all other cases. First, as stated in 1.5, one of the aims of the study was to evaluate the use of predominately machine derived indices. Such methods of obtaining information have the benefit of being consistent in terms of the values derived for each index and have a practical benefit as they are relatively less time-consuming than organising judges' time. Second, there are significant issues when involving judges which require careful consideration, but that would remain a source of uncertainty even after operationalisation. For example, striking a balance between training judges and utilising their existing expertise effectively is a challenge and some level of uncertainty that a suitable balance was achieved would always likely to remain. According to Alderson (2000), training would be expected to improve agreement between judges but risks creating a 'cloning' exercise, which prioritises conformity with the researchers' conception of what is to be judged over the accumulated expertise of the independent judges.

The best way to mitigate uncertainty in judging exercises, is, perhaps through careful matching of the expertise the judges possess to the judgement task they are to be given (Alderson & Kremmel, 2013). For example, linguists would be best placed to determine linguistic characteristics of the text but it would be less appropriate to ask them about the relative difficulty of certain lexis for a candidate of a particular ability level. This, on the other hand, would be a more suitable question for a language teacher. The appropriate matching of judge to task is likely to ensure enhanced accuracy, but also mean that they would require little or no training, thus mitigating concerns about cloning (Alderson 2000). This was indeed the approach taken in the current study and details are provided below. Before further discussion concerning the judges can take place, however, some consideration of the required judgements is needed. If machine-based indices are to be used instead of direct judgements, text must still be processed. For these indices, therefore, the focus is on to which text this should be.

An entire reading passage could be analysed to produce indices for a study. However, where more than one item pertains to the text, variables would be the same for each item. In this case, it would be impossible to relate difficulty to attributes. Freedle and Kostin (1993) chose to base their indices on the whole text. This was no doubt related to the scale of their study, which comprised 100 reading passages and 213 items, categorised into three item types: main idea, inference, supporting idea. They therefore removed items so that no passage contained more than one item of any one item type. Their study also used a subset of this data which comprised no more than one item of any type per passage, which ensured that variables did not replicate others. Variables were then derived from the text relevant to each item.

As with Freedle and Kostin (1993), Embretson and Wetzel (1987) reduced the number of items analysed from their initial 12 test forms (six each from two different tests). For their data, they stipulated that no more than one item per paragraph was to be included, although it is not explained whether the selection

process depended on a judgement of whether an item related to a particular paragraph, or whether it was stipulated in the item stem or rubric. Whichever method was used, the process yielded 29 items from one test and 46 from the other.

Both Buck et al. (1997) and Aryadoust and Goh (2014) employed judgement to determine the *necessary information* (NI) for each item contained in the relevant passage. For Buck et al. (1997), after the NI was selected, judgment was used to construct contextual variables based on the characteristics of the NI for each item. Agreement indices (percentage of agreement, or Pearson correlation coefficients, depending on the number of levels in the variable) were then used to compare the resulting variables and these indices were used to appraise the success of their efforts. They report high levels of agreement for these indices: from 73% to 99%.

By contrast to Buck et al. (1997), although Aryadoust and Goh (2014) used the NI they found for each item, they combined it with the text for the stem and options before processing it using Coh-Metrix (McNamara et al., 2012) to generate the indices required for their main analysis. Human judgement was not involved after this, as Coh-Metrix converted the text into indices, so agreement indices would only be applicable to judgements concerning the NI. Aryadoust and Goh (2014), however, did not report any agreement indices. This may be because calculating such indices is difficult when judgements are not made with reference to a limited number of categories.

When using judgement to construct variables directly for items, such a scale might be binary (e.g. present, absent), or consist of several levels concerning the relative presence of the attribute in the item (e.g. absent, low, medium, high). However, when judgement involves selecting segments (e.g. sentences) of text from a reading passage, the number of possible categories is very large (every sentence in the passage). According to Gwetl (2012), agreement indices which account for chance judgements can either be based on contingency tables which include cells for each combination of possible categories which a judge could select, or internal

consistency indices, like Cronbach's Alpha. These indices require a scale consisting of all possible judgements to be coded into the data before they can be applied. If agreement indices do not account for chance judgements, simple indices such as the proportion of agreement in all judgements made can be calculated (Hulstijn, 2014).

The current study involved the analysis of a single test version, so items could not be removed without adversely affecting the subject of study. For this reason, expert judgement was used to determine the parts of the text most relevant to each item, like Buck et al.'s (1997) NI; in this thesis it is referred to as the *relevant text* for an item. For most variables, the relevant text was used as an input for the construction of variables by machine. The nature of the expertise required for judgements in the current study was that of items writers and editors, as they are familiar with considering the way in which items function and how the stem relates to different parts of the reading text. It was also seen as appropriate to calculate simple agreement indices to determine the veracity of the judgement process.

2.9.3 Main analysis methodology

2.9.3.1 Mathematical models in experimental for componential analysis

Among the methods Messick (1989) listed for investigating cognitive processes were, mathematical modelling and psychometric modelling. Mathematical modelling seeks to decompose the sources of difficulty in tasks by specifying the relationships between them in the form of equations and applying the model to data (Embretson, 1983, 1985; Sternberg, 1985). Alternative models may be fitted to data and tests of model fit compared in order to determine which provides a better substantive explanation of the data (Embretson, 1983). The initial approach developed by Sternberg (1985) involved eight steps. The first three steps deal with decomposing a complex cognitive process into subtask components according to theory. Data is gathered on candidate performance on each component, usually through specifically designed instruments, administered to respondents individually in an experimental setting. These initial steps are much like the

approach of Carr et al. (1990) to developing a model of reading comprehension with a battery of 15 tests. A mathematical model (typically a regression model) is then developed (step 4) to specify how the components relate to each other to during performance of the complex task. Various stages of testing follow to ascertain the extent to which the data support the formulated model.

Table 12 Steps in componential analysis (Sternberg, 1985)

	Step
1	Selecting or generating a theory of relevant cognition
2	Selecting one or more tasks for analysis
3	Decomposing task performance
4	Quantification of componential model
5	model Testing: initial validation
6	model Testing: external validation
7	Reformulation of componential model
8	Generalisation of componential model

The componential analysis of Sternberg (1985) and Carr et al. (1990) is suited to studies where an experimental approach may be adopted. This is not usually the case where administering a large test battery is considered logistically too challenging, too intrusive, or where a live test is being investigated, as in the current study. Mathematical regression models may still be used under these circumstances, although the distinction between components of the cognitive process is more difficult.

2.9.3.2 Other uses of mathematical models

2.9.3.2.1 Multiple linear regression

Freedle and Kostin (1993) employ multiple linear regression to determine the relationship between item attributes and item difficulties. In their study, 213 items from 20 reading tests were used, with responses from around 2,000 candidates in total. The data set was partially crossed (some candidates responded to all items), so the data were linked and the relative difficulty of each item could be calculated based on the results. Item difficulty was one of the variables used as a dependent variable in the subsequent regression. As Buck et al. (1997) point out, multiple

regression, among other things, requires variables with a large number of cases to produce results with sufficient power for useful conclusions. Freedle and Kostin's (1993), used 213 items from 20 different test forms. Smaller studies focussing on a single test, with perhaps 30 items, will produce results with far more measurement error, which will, in turn, make it difficult to draw sound conclusions. A further disadvantage of linear regression is that is assumes unidimensionality and therefore may be unsuitable for some data, as it would lead to erroneous results (Doran, Bates, Bliese, & Dowling, 2007).

2.9.3.2.2 Tree-based regression (TBR)

Another type of regression employed in studies which seek to decompose test difficulty is tree-based regression (TBR). Gao and Rogers (2011), Rupp et al. (2001) and Aryadoust and Goh (2014) use TBR to investigate cognitive properties underlying responses to test items by specifying IRT (Item Response Theory) item difficulty parameters as the dependent variable. The independent variables are contained in an incidence matrix. The analysis forms a tree-like structure, consisting of nodes representing groups independent variables, progressively split into smaller nodes which form successive levels and are linked by branches to their parent node (hence 'tree-based regression'). The terminal nodes display the final groupings against the dependent variable, the difficulty scale. The technique can use a mixture of categorical and scale variables and does not make assumptions about distribution of the variables (Gao & Rogers, 2011).

Because Gao and Rogers (2011), Rupp et al. (2001) and Aryadoust and Goh (2014) specify IRT-calculated item difficulty estimates as the dependent variable, the approach is pseudo-psychometric and is likely to obtain similar results to a fully psychometric approach like Cognitive Psychometric Models (CPM) (see 2.9.3.3) (De Ayala, 2009). Because measurement error for the dependent variable is not taken account of in TBR, and TBR does not estimate measurement error for any of its parameters, the significance of parameter values is not as robust as for Cognitive Psychological Models (Embretson & Reise, 2000). In other words, the value of the findings is shrouded in greater uncertainty.

Two further issues should be considered when evaluating the utility of TBR. The first is the dimensionality of the data, as like multiple linear regression, unidimensionality is assumed. The second is the need to use training data before conducting the main explanatory analysis (Aryadoust & Goh, 2014). TBR requires training data, which is used to improve estimation before the data of interest is analysed. As its influence on the main analysis will be considerable, training data should be as close as possible to study data (Aryadoust & Goh, 2014). In order to obtain suitable training data, some of the study data is usually sacrificed, and it can no longer be used for study. For the current study, this would be a particular concern, as the aim is to investigate the items of a single test which contains four different tasks, each with different tests and item types.

2.9.3.3 Item Response Theory (IRT) and Cognitive Psychometric Models (CPMs)

An alternative to mathematical modelling may be found in CPMs, which model the response to an item as the interaction between properties of the item (difficulty) and properties of the candidate (ability). These models are from the IRT or Rasch family of models but have been extended to examine cognitive processing (Gorin & Svetina, 2012). Two principal advantages of these models are that they i) control for variation of difficulty or ability in other item parameters, such as those related to estimating item attributes, and ii) parameters are estimated on a readily interpretable scale which is based on the likelihood of success on items (Embretson & Reise, 2000). Regression analysis with pre-calculated item difficulty as the dependent variable also has these advantages, although estimates of error are likely to be less accurate (Embretson & Reise, 2000). CPMs, as will be shown in 2.9.3.3.4, also have the advantage, depending on implementation, of far greater flexibility in the way data can be modelled. For example, dependency between items, a violation of the assumptions of IRT models can be modelled and, therefore, controlled for (De Boeck et al., 2011; Tuerlinckx & De Boeck, 2004). The Rasch model is also outlined, as it is an archetype of psychometric models, and the basis for the CPMs discussed.

2.9.3.3.1 Direct modelling of cognitive components using CPMs

The direct modelling of components has been proposed by some researchers, where they are treated as dimensions. Whitely (1980), for example, introduced a multidimensional CPM, with estimated parameters for each component (Multicomponent Latent Trait Model – MLTM). Components were thus modelled as latent traits representing candidate performance on each component of the complex task. Parameters for attributes were later added to the model so that the influence of contextual effects could be modelled and the General Multicomponent Latent Trait Model (GLTM) was produced (Embretson, 1984). These models were problematic to implement, however, mainly because software to do so was not readily available and it was found to be difficult to recover parameters from the model. Embretson and Yang (2006) published code to allow both MLTM and GLTM to run on SAS (SAS Institute Inc., SAS) and Bolt and Lall (2003) produced a WinBugs (Lunn, Thomas, Best, & Spiegelhalter, 2000) implementation of MLTM. In the case of the latter, parameter recovery was found to be problematic when applied to reading test data. Whitely (1980) and Embretson (1984), however, were able to implement MLTM and GLTM on data from verbal analogy tests, but the components and attributes were far more distinct than for Bolt and Lall's (2003) data, and therefore easier to estimate.

More recently, Embretson and Yang (2013) have sought to overcome problems with earlier models by introducing the Multicomponent Latent Trait Model for Diagnosis (MLTM-D). Unlike MLTM and GLTM, it is not exploratory in nature and the nested structure of the data (items, attributes and components) must be comprehensively specified in advance. Embretson and Yang (2013) provide guidance on ensuring that the model is *identified*. Identification is a technical concept important to mathematical and psychometric models, indicating whether the parameters estimated by the model are unique, or, whether, given the data, another set of values could have been estimated with equal probability. If the values are not unique, it would mean that the researcher is no nearer knowing the answer to their research questions, as alternative results might equally be true.

Such situations typically come about if the amount of information to be estimated is larger than the amount of information in the data but could also be caused by mode complex reasons which are harder to determine in advance (Kenny & Milan, 2012). MLTM-D requires, among other things, that some parameters are fixed and that not all items require all components to be completed successfully. These requirements mean that the model may be suitable for complex tasks where components are easily separable, both conceptually and practically, as in the worked example of a mathematical achievement test given by Embretson and Yang (2013). For more integrated tasks, such as those involved in reading, where components and attributes in the composite model (see 2.8.1) are not easily separable, the model is unsuitable.

2.9.3.3.2 The Rasch model

Although it would be desirable to model the influence of cognitive components in the study data, such a step would be unlikely to be successful for the reasons given in 2.9.3.3.1. It is however, still possible to model the contextual effects using other CPMs, specifically the Latent Logistic Test Model (LLTM). A more in-depth understanding of CPMs is first required, however, and this will be given by explaining the Rasch model, which is not a CPM, but from which the LLTM was derived.

The Rasch model, also known as the one-parameter IRT model, is given in Equation 1. The key feature of the model to note is that likelihood of scoring 1 on an item is directly related to the difference between candidate ability and item difficulty. If the level of ability is higher than the item is difficult, the chances of success are greater than 0.5, if it is lower, chances of success are below 0.5. Application of the model to data places items and candidates on the same scale and thereby defines the latent trait. A more advanced position on the latent trait equates to a higher ability for candidates, or more difficulty for items. In other words, more advanced candidates have higher likelihood of success on all items than less advanced candidates; more advanced items yield a lower likelihood of success for all

candidates than less advanced items (De Ayala, 2009; Embretson & Reise, 2000; Wilson, 2005).

$$P(x_{ij} = 1|\theta_j, b_i) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}}$$

where

$P(x_{ij} = 1|\theta_j, b_i)$ is the likelihood of obtaining a score of 1 conditional on θ_j and b_i.

θ_j is the ability of candidate j

and

b_i is the difficulty of item i.

Equation 1 The Rasch model (Rasch, 1980)

The Rasch model, and many other IRT models, have two main assumptions:

- unidimensionality
- local item independence (violations of which are termed *LD*)

The former requires that all items tap the same underlying trait, the latter that the response to one item does not depend on the response to any other item (De Ayala, 2009), although similar violations can result directly from contextual effects influencing more than one item in a similar way, and thus inducing dependency between them, rather than directly from other items (Wainer, Bradlow, & Wang, 2007). Unidimensionality is important because item difficulty and person ability is estimated on a single measurement scale. If the data contain more dimensions (i.e. more than one ability is tested), estimates will be a composite of the position of the item or person on all the relevant dimensions. In the case of multidimensional data, a multidimensional model would provide estimates for the position of each item or person on each dimension. In this way, the structure of the data, and the abilities it relates to would be more accurately represented. In the case of such a model, the assumption would be, rather than pure

unidimensionality, that the dimensions specified in the model would adequately account for the structure of the data.

LD can result from many possible causes, such as a correct response to one item implying the key to another item or dependency built in to the response format (Lee, 2004), as with multiple matching tasks, where it is possible to wrongly select the key for one item as a response to another, and thus ensure that two items are answered incorrectly. Wainer et al. (2007) also discuss effects such as speededness, which involve unmodelled features of the context, here the task setting, influencing the probability of success. Although this situation does not indicate the influence of one item over another, such items are not statistically independent because of a common but unmodelled influence. That LD is problematic may be seen from Equation 1, where the probability of a particular candidate's success on a particular item is given as the ability of the candidate minus the difficulty of the item. A term for the influence of other items on the probability of success is not included, so any such influence is pooled into measurement error.

LD often has the effect of creating additional dimensions in the data (Thissen & Steinberg, 2010). This is because the more the responses to one item are influenced by another, or the responses to two or more items are influenced by unmodelled effects, the more response patterns co-vary. Covariance is the basis of dimensions in data because it reflects similarity (Child, 2006). Considered in another way, a violation of the unidimensionality assumption is also a violation of the local item independence assumption because in both cases, similarity between items is unmodelled (Andrich & Kreiner, 2010). Because LD is linked to the existence of dimensions, multidimensional models have been developed to account for LD. For example Testlet Response Theory (Wainer et al., 2007) and bifactor models (Reise, 2012) subsume contextual effects into dimensions.

In the case of both Rasch model assumptions, it is difficult to quantify the importance of any violations. Completely unidimensional data would mean that all items are identical (Reckase, 2009), which would be of no value. Furthermore,

since local item independence implies items which are statistically independent of each other, no covariance should exist between items, which would imply each requires its own dimension. It is important to find a balance, therefore, between the assumptions and the use being made of the data (De Ayala, 2009). Smith (1996:27-8) argues for *practical or functional unidimensionality* as a departure from *theoretical unidimensionality*, whereby some multidimensionality may be tolerated but the amounts depend on the use of the test. In a simulation study examining the robustness of IRT models to violations of unidimensionality, Albano (2014) examined data representing a number of different experimental conditions: the number of items, the number of dimensions, the balance of items per dimension and correlations between dimensions. Under all conditions, parameters were found robust to dimensionality, although data with fewer items, more dimensions, a more unbalanced distribution of items per dimension and lower inter-correlations were responsible for more inaccurate results. As a consequence, he recommended equalising the influence of additional dimensions where possible. Measures to detect and address dimensionality and dependency will be discussed in 2.9.3.3.4.

Wilson and Moore (2011) and De Boeck and Wilson (2004a) divide IRT models into descriptive models, which are used for measurement of persons, items or both, and explanatory models, for investigating properties of persons, items, or both. They classify the Rasch model as doubly descriptive, as it estimates both person ability and item difficulty, according to their framework (Table 13). For the current study, an item explanatory model was required, which will be discussed next in 2.9.3.3.3.

Table 13 Models as a function of the predictors (De Boeck & Wilson, 2004a:47)

Item predictors	Person predictors	
	Absence of properties	**Inclusion of properties (person properties)**
Absence of properties	Doubly descriptive	Person explanatory
Inclusion of properties (item properties)	Item explanatory	Doubly explanatory

2.9.3.3.3 Linear Logistic Test Model (LLTM)

Fischer's (1973) Linear Logistic Test Model (LLTM) is an item explanatory model, adapted from the Rasch model (2.9.3.3.2). LLTM is given in Equation 2 and contains only one significant difference from the equation for the Rasch model (Equation 1), which is that the item difficulty parameters have been replaced by estimates for attribute difficulty and their regression weights. According to this model, likelihood of success on an item depends on the ability of the candidate and the sum of difficulty of the attributes of that item, as specified by the incidence matrix. Explanatory models like the LLTM are not expected to provide accurate measurement in the same way as descriptive models (De Boeck & Wilson, 2004a) but, by the same token, descriptive models provide little useful information about cognitive processes (Gorin & Svetina, 2012).

$$P(x_{ij} = 1|\theta_j, \eta_m) = \frac{e^{\theta_j - (\sum_m c_{im}\eta_m)}}{1 + e^{\theta_j - (\sum_m c_{im}\eta_m)}}$$

where

$P(x_{ij} = 1|\theta_j, \eta_m)$ is the likelihood of obtaining a score of 1, conditional on θ_j, η_m, d

θ_j is the ability of candidate j

c_{im} is the difficulty of attribute m on item i

η_m is a regression weight for attribute m

Equation 2 The Linear Logistic Test Model (Fischer, 1973)

The LLTM carries the same assumptions as the Rasch model and Fischer (1995) recommends that these are first tested by fitting the Rasch model to the same data and examining the assumptions based on the results. In addition, according to Rijmen and De Boeck (2002), each attribute is considered to affect each person's probability of success equally (the effect of the attribute is not different for different candidates) and the attributes are not tapping the same trait (they are multidimensional).

2.9.3.3.4 Random Weights LLTM (RWLLTM)

Rijmen and De Boeck (2002) developed a multidimensional CPM by generalising the LLTM to produce the Random Weight Linear Logistic Test Model (RWLLTM). A primary motivation was the assumption of the LLTM that attributes affected candidates to the same extent (2.9.3.3.3). In order to relax this assumption, they adapted LLTM so that some attributes were allowed to vary over persons, which has the effect of creating a dimension for each level of a factor-based variable for which candidate abilities are estimated. Together with LLTM, this model has been adopted for the current study for reasons which will be explained below. Before giving the reasons for this, however, the model will be described, and this can be more easily explained within a Generalized Linear Mixed Models (GLMM) framework (De Boeck et al., 2011; De Boeck & Wilson, 2004b). This framework is preferred, as it allows great flexibility when adding parameters to a model (De Boeck & Wilson, 2004b). The equation representing the model, reproduced in Equation 3, is provided by Rijmen and De Boeck (2002) as a formula which evaluates the logit probability of success on an item, which is commonly done when discussing such models in a GLMM framework.

$$logit[P(X_{in} = 1|\mathrm{A}, \eta, \lambda, \theta)] = \theta_n + \sum_{p=1}^{P_1} \alpha_{ip}\eta_p + \sum_{p=P_1+1}^{P} \alpha_{ip}\lambda_{np}$$

where

P_1 is the number of fixed effects

P is the number of fixed plus random effects

$X_{in} = 1$ if the n^{th} person responds successfully to the i^{th} item

A the incidence matrix which contains the attribute information

α_{ip} the attribute for the i^{th} item and p^{th} person

η the fixed effect parameters

η_p parameter for the p^{th} fixed effect

λ_n parameter for the n^{th} random effect

θ_n ability for the n^{th} person (random effect)

Equation 3 RWLLTM (Rijmen & De Boeck, 2002:274)

GLMMs are conceptually very similar to (multiple) linear regression models: a dependent variable is predicted by one or more dependent variables (Stroup, 2013). In terms of the Rasch model (2.9.3.3.2), this can be understood as the score a candidate obtains on an item being predicted by a combination of the item's difficulty and the candidate's ability. As item difficulty is replaced by the sum of the difficulty of the modelled attributes in the LLTM, this model has a form like multiple regression equation where the item attributes are independent variables (Equation 3). Unlike linear regression, and like the Rasch model, the relationship between the dependent and independent variables in psychometric models is probabilistic. This is specified in a GLMM through a logit 'link function'. The link

function connects the linear component, which consists of the independent variables, with the dependent variable (the score) in order to estimate the probability of obtaining the score for a given candidate-item combination. The linear component comprises what are called the *fixed effects*, which correspond to regression weights in a regression model. It is the estimation of these effects which are of interest to the researcher. In essence, they are estimates of the effect of independent variables (attributes or item difficulty) on the dependent variable (score). One further element, which makes the model a mixed model, is the inclusion of a random component. 'Random' in this context means variance which the model does not attempt to explain through a deterministic relationship to other elements, as fixed effects are (Stroup, 2013). They are used to explain the variance of elements which are usually not of direct interest to the researcher and are therefore estimated for each level of categorical variables (such as with person ability estimates). In sum, the fixed effects summarise the impact of the variables of interest, and the random effects may be seen as controlling for the impact of nuisance effects (Rijmen & De Boeck, 2002).

Unlike the models described in 2.9.3.3.1, the RWLLTM does not model the cognitive components directly. As with Embretson and Wetzel (1987), the approach adopted in the current study was to define attributes according to an understanding of the components but to model the attributes without the components. This is reasonable, given the importance of all components to each item, and the interrelationships between the components (see 2.8.1). In other words, the RWLLTM has the advantage that, contrary to the MLTM-D, model identification does not require some items to be excluded from some dimensions. Instead, requirements concern the number of items compared to the number of model effects and the potential to use matrix algebra with the matrices containing the data. Rijmen and De Boeck (2002) list them as:

- that there should be more items than the sum of fixed and random effects

- that the response data matrix, the matrix containing the attribute variables and the combination of the two should all be of full column rank[6]
- that the covariance matrix for the random effects must be symmetric positive definite[7]

Assumptions are the same as for LLTM (2.9.3.3.3) and, within a GLMM framework, the assumption that attributes tap distinct traits may be considered as absence of high *colinearity* (or correlational relationships) between them. According to De Boeck et al. (2011), the impact of violations of this assumptions are limited: the estimates for the fixed effects may be misleading but estimates to the model as a whole would still be valid. A final assumption is common for GLMMs: that the distribution of the person parameters is assumed to be normal.

The data for this study were derived from a test with four different tasks, each with a different test method and a shared reading passage. The RWLLTM was therefore expected to be useful because of the possibility to specify a dimension for each task to account for contextual effects which relate to the tasks but not accounted for in other ways. In other words, items within tasks were expected to co-vary more with each other than with items from other tasks. The effect of adding this dimension would be that candidate abilities are estimated separately for each task. Accounting for 'nuisance' variance in this way would decrease model error and allow more accurate estimation of fixed effects. The use of RWLLTM would, however, depend on violations of either the unidimensionality assumption, local item independence, or both. In addition to specifying random weights, parameters to control for remaining instances of LD could also be added to the model (De Boeck et al., 2011). In the case where a random weight is specified for each test task, rather than an expectation of unidimensionality, there would be

[6] a technical consideration that specifies that columns in the matrix cannot be derived using linear equations from other columns in the same matrix

[7] another technical consideration whereby when non-zero vectors of the matrix are multiplied by their transpose, the result is greater than zero

the assumption that the specified structure would adequately account for the structure of the data.

A summary of the advantages and disadvantages of the analytical models considered for the current research is presented in Table 14. As discussed above, RWLLTM is the most suitable model and would be adopted if the use of random weights offered a significant improvement over the simple LLTM. The method for testing the difference will be described in 3.8.2.1. Henceforward, the proposed model will be referred to as 'LLTM' even if it contains random weights. This is because the addition of random weights is one of two types of modification which may be applied to the LLTM in this study. The other is described in the preceding paragraph and concerns LD. As a result, the use of either correction is considered a modification of LLTM, rather than a different model.

Table 14 Comparison of analytical methods for construct representation

	Regression		CPMs				
	Linear multiple regression	**Tree-based regression (TBR)**	**MLTM**	**GLTM**	**MTM-D**	**LLTM (GLMM implementation)**	**RWLLTM (GLMM implementation)**
Models components	No	No	Yes	Yes	Yes	No	No
Models attributes	Yes	Yes	No	Yes	Yes	Yes	Yes
Unidimensionality assumed	Yes	Yes	No	No	No	Yes	No
LD assumed	Yes	No	Yes	Yes	Yes	No	No
Parameter recovery	Easy	Easy	Difficult	Difficult	Unknown	Easy	Easy
Identification	Unproblematic	Unproblematic	Problematic	Problematic	Problematic	Unproblematic	Unproblematic
Error estimated for parameters	No	No	Yes	Yes	Yes	Yes	Yes
Estimation speed	Fast	Unknown	Unknown	Unknown	Unknown	Slow	Slow
Training data required	No	Yes	No	No	No	No	No
Large number of cases required	Yes	No	No	No	No	No	No
Software widely available	Yes	Yes	No	No	No	Yes	Yes

2.10 Research Questions

In order to investigate the construct representation of FCE Dec 2005, the attributes specified in 2.7 were analysed using the LLTM (2.9.3.3). The primary results were the estimates of fixed effects for each of the attribute indicators (variables). The first research question below is relevant to these results. Since these attributes were considered as nested in subcomponents and components, the second and third research questions were included to investigate them. As discussed in 2.8, the findings for the subcomponents and components were only partial, so a full understanding of each was not to be expected as a result of the current research. The penultimate research question was of interest because contextual features connected with the test method are likely to have a significant effect on the cognitive processes of test candidates. Findings for this question were likely to come from the OP and RD components in particular. By contrast, it might be expected that the contextual effects in SEARCH and READ would be the same regardless of differences in task. The fifth research question was designed to gain an understanding of how well the model described the data, given the particular circumstances of this research. As stated among the aims of this research (1.7), a practical method to investigate the construct representation of single test forms was of interest.

1. Which contextual attributes (see 2.7) can be shown to influence the difficulty of FCE Dec 2005, and by how much?
2. Which subcomponents included in the composite model (see 2.7) can be shown to influence the difficulty of FCE Dec 2005, and by how much?
3. Which components included in the composite model (see 2.7) can be shown to influence the difficulty of FCE Dec 2005, and by how much?
4. What evidence can be found of test methods effects influencing item difficulty?
5. What proportion of the variance of the corresponding Rasch model does the LLTM account for?

2.11 Chapter summary

In this chapter, three theoretical models concerning reading were described and a composite mode formed for the purpose of investigating the construct representation of a reading test. The model consisted of four components: OP, SEARCH, READ and RD. For each component, a number of attributes were posited, based on theory. A psychometric model was presented with which to investigate reading tests (LLTM) and research questions proposed.

3 Method

3.1 Introduction

This chapter contains a description of the method adopted for this study. After describing the data, the method used to prepare the response data for later analysis and the production of descriptive statistics is described. Based on linguistic and other features of the test materials, the construction of indicators, or variables, for the main analysis is described. Finally, the fitting and modification of a number of models, tests of their comparative fit and quantification of variance explained is detailed.

As discussed in 1.7, the aims of the analysis were to:

- determine elements of the construct representation of the Reading paper of a form of First Certificate in English (FCE) administered in December 2005 (FCE Dec 2005)
- develop a practical method which can be deployed in the construct investigation of reading tests with varying test methods
- trial the use of machine generated indices in the construct investigation of reading tests

The analytical method adopted for this purpose was a Generalised Linear Mixed Model (GLMM) implementation of the LLTM (see 2.9.3.3.3, 2.9.3.3.4). In such a model, coefficients are estimated for each feature of interest, using response data and an incidence matrix containing variables, or indicators[8] as they are referred to in such models, which provide information about items. The indicators include counts such as the number of syllables, categories such as whether the order of items follow the order parts of the reading passage to which they refer (i.e. yes/no)

[8] The term 'indicator' will be used to refer to a variable which is used in a GLMM.

and continuous indicators such as indices representing the level of hypernymy in an item.

Before the main analyses were completed, however, various descriptive statistics were calculated. After describing the data, this chapter will begin by describing them. The purpose of these descriptive statistics was to present a picture of the data which may aid interpretation of other statistics at later stages. In particular, it was important to ascertain i) the resemblance of the sample taken from the data to the original data set, and ii) the resemblance of the materials to those of other forms of the same test. The former ensured that the sampling process did not have any unintended effects, and the latter acted as supporting evidence for the generalizability of the findings. In addition, in some cases, it indicates that the assumptions of later analyses are met.

3.2 Description of the data and materials provided

3.2.1 Response data

The data were provided by Cambridge ESOL, the exam owner, in the form of a .txt document containing a person by item response matrix. Responses were in the form of a letter, indicating candidate choice for each item (column), and 'O' representing missing data. A candidate identification number occupied the first nine characters of each row (candidate) of the matrix. For each column (item), the key, the number of options per item and the number of the task the item belonged to were also provided (see Figure 6).

```
35 O O 09
EBHDAGCDABACBDBCHAEBGFCACABABDACBDC
99999994444444499999999999999999999
11111111111111111111111111111111111
IT2650001EGHDBFCABCABBCDCAFBDEHADDCBABCDCDAC
IT2650002EBHFAGCDABBCCDBCHAEBGFBACABDBDACDDC
IT2650003EFHBAGCDCDACBDADAFEBGCAACCBABDDCDDC
IT2650004EFHDBGCABABABADFAGECBFDACADABCDCBDC
IT2650005EFDBAGCBCDCCDADCEADBFHCDOABCCDOBADA
```

Figure 6 A sample of response data from FCE Dec 05

3.2.2 Candidate background characteristics

Data concerning background characteristics of candidates was provided in an Excel Workbook, arranged as a person by question matrix. It was possible to cross-reference the data in the Workbook with the response data through the common candidate identification number. The candidate background information is routinely collected by the test provider as part of the exam administration process. Candidates are asked to record their responses to questions contained in the Candidate Information Sheet (CIS). An example dated from 2006 (Hulstijn, 2011), which is expected to be very similar, is contained in Appendix 3: candidate background information form. All variables are categorical, with candidates being asked to select from amongst a range of options. Those variables for which data was provided are listed in Table 15.

Table 15 Variables contained in the CIS data

	Variable	Response categories
1	Age	<10, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26-30, 31-40, 41-50. 51+
2	Gender	Female, Male
3	L1	Any one of 77 listed languages
4	Nationality	Any one of 150 listed nationalities
5	Previously attempted this exam	No, Yes, Yes – more than once
6	Attended exam preparation classes	No, Yes at language school, Yes at college, Yes at work
7	Educational level	Primary School, Secondary School, College or University

3.2.3 Test materials

Facsimiles of the relevant test materials were provided to the researcher in electronic form as a Word (.doc) document. These were formatted as they appeared on the exam papers when administered to the FCE candidates and are available in Appendix 1: test papers.

3.3 Crossing, cleaning and preparation of the data and materials provided

3.3.1 Response data

3.3.1.1 Preparation

3.3.1.1.1 Data crossing

A fully-crossed data set includes information in each cell of the matrix. A more crossed data set was considered to be better for the main analysis, as GLMM requires matrix algebra. Conducting such analysis with sparse matrices can mean that it is impossible to complete the analysis. For the current study, in order to produce a more crossed data set, large sequences of missing data (e.g. more than five consecutive missing responses) were removed on a casewise basis using Excel (Microsoft Corp., 2010a). In other words, all data pertaining to such candidates (the case) were removed, but data related to the item (the variable) from other candidates was retained. In total, this accounted for the removal of 165 candidates.

3.3.1.1.2 Data cleaning

The aim of cleaning the data was to remove data which was likely to be erroneous and could not be used in later analysis. This primarily consisted of removing responses which, other than those coded as missing, were not appropriate responses to the given task. Such responses consisted principally of letters higher than H for Parts 1 and 3, and higher than D for Parts 2 and 4. This process was carried out using Excel (Microsoft Corp., 2010a) and 2,080 candidates were removed from the data as a consequence. The crossed, cleaned data set therefore contained a total of 25,803 candidates.

3.3.1.1.3 Creation of score matrix

An additional person by item matrix containing candidate scores was derived from the response data matrix and the key. The letter 'O' was retained for missing values.

3.3.2 Preparation of the response data for further analysis

3.3.2.1 Rationale

It was decided to reduce the size of the data to be used in the remaining analyses for two reasons. First to ensure that the data was not skewed by large numbers of responses of candidates from a small number of L1 groups, and second, to reduce

processing time when running the analysis, which is considerable for GLMM[9]. A sample of around 10,000 cases was therefore targeted.

3.3.2.2 Construction of sample

The size of the final random stratified sample was 9,961 cases. It was taken from the cleaned crossed data as follows. Stratification by L1 was used in order to ensure the sample was representative of the initial data set but that no particular L1s dominated the sample. To this end, L1s with more than 900 cases in the original data set were rendered approximately equal in number by random sampling. An additional group of all other L1s was retained unchanged. To obtain the sample, the response file was first augmented with candidate background information. The data were divided into subsets according to the L1 groups as outlined and the data for each group imported into SPSS (IBM Corp., 2013) using the random sample feature, which allows the user to specify the approximate percentage required. These sampled subsets of the data were then recompiled into a single data set.

3.3.2.3 Validation of sample

It was important to verify that the sampling did not distort the data. For this reason, a comparison between the original response data and the sampled data was necessary. To do this, frequency tables were constructed from the raw response data and the score data. These tables are available in Appendix 5: summary of response matrices and Appendix 6: summary of the score matrices. Calculations were done using Excel (Microsoft Corp., 2010a) and percentages were used to facilitate original data to sample comparison. An Independent-Samples Wald-Wolfowitz Runs Test (Wald and Wolfowitz, 1940) was also conducted, comparing the scores on each item in the sample data set to those in the crossed, cleaned data set. This test was chosen because the score data was binary and

[9] Some analyses lasted more than ten hours. If the original data set were used (approximately 30,000 candidates, or three times the sample), the result would not be an analysis lasting 30 (10 hours times three) but considerable more. This is because matrix algebra is involved and the number of calculations does not increase linearly. Taking a sample can therefore be considered an important step if the method employed here were to be replicated within a test development and administration cycle.

ordinal. A test such as the Independent Samples T-test is unsuitable because it assumes a normal distribution, which cannot be the case with binary data (Mislevy, 1984). The Wald-Wolfowitz Runs Test involves comparing sequences, or 'runs', of numbers within both samples. It acts as a significance test where the null hypothesis is that the distribution of the variable is the same across all categories. Results from the test are summarised in 4.2.1, with full results in Appendix 4: Independent-Samples Wald-Wolfowitz Runs Test results.

Since the sample was stratified by L1 to ensure that it was representative of the original data set, it was felt important to verify this by comparing the frequency of relevant background characteristics from the original data and the sample. To do this, the same descriptive analyses conducted for the original sample was repeated with the sample data and the analyses were then visually compared to one another. In the case of most statistics, the expectation was that the results would be highly similar for both the original crossed and cleaned data and the sample. However, since the sample was stratified by L1 for the largest L1 groups, these L1s were expected to appear with around equal frequency to each other. Furthermore, other variables related to L1, for example, nationality, would also be influenced by the sampling procedure. The results of this analysis may be found in full in 4.2.2.2 and Appendix 8: descriptive statistics for candidate background data.

In addition to the statistics for the original crossed, cleaned data set and the sample being presented together, comparable data given by Khalifa and Weir (2009), derived from candidates who sat the test in 2007, is included. This was done to determine how representative the sample is of the test in general. If other test forms yield similar data, the December 2005 form can be considered typical and conclusions drawn from the main analysis are therefore more likely to be generalisable. For this reason, the analysis was conducted and summarised so that it could be compared to the general overview of FCE Reading presented by Khalifa and Weir (2009). As Khalifa and Weir only present data on a subset of candidate background variables, and as not all of the variables included on the CIS form were

made available by the exam owner for the current study, only L1, age, gender, educational level and attendance of preparation classes were included.

3.4 Description of the data and materials

3.4.1 Descriptive statistics for the response and score matrices

In order to obtain a better understanding of the data, a number of classical statistics and indices were calculated from the scored response data for each task and for the data overall using the Classic software package (Jones, 1998). The resulting analysis may be found in 4.2.2.1 and Appendix 7.

Table 16 Descriptive statistics generated for the response data after crossing and cleaning

	Statistic or Index
1	Mean score
2	Median score
3	Modal score
4	Variance
5	Standard Deviation
6	Skew
7	Kurtosis
8	Cronbach's Alpha
9	Standard Error of Measurement (SEM)
10	Mean P
11	Mean Item-total
12	Mean Biserial Correlation
13	Frequency, Cumulative Frequency (as a raw count and as a percentage) for each point on the score scale

3.4.2 Analysis of task texts

Analysis of task texts was conducted for two reasons, to allow comparison with a reference set of statistics obtained from Khalifa and Weir (2009:76, 122, 131) and to provide a broad summary of the characteristics of the texts in the study. The former reason is important because it would help to determine the extent to which findings from the study may be generalised to other forms of the same test. The statistics described in this section, therefore, compliment those in 3.3.2.3 which were matched to corresponding statistics presented by Khalifa and Weir (2009). The variables for comparison were:

1. Overall number of words
2. Mean words per sentence
3. Flesch reading ease
4. Flesch-Kincaid grade level
5. Tokens
6. Types
7. Type-token ratio
8. Tokens per type
9. K1 words
10. K2 words
11. AWL words
12. Off AWL list words
13. Lexical density

In Khalifa and Weir (2009), the figure for text variable 1 was the target total text length and was not based on empirical research into the number of words observed in the texts used on the test. Those for 2 to 4 were derived by Khalifa and Weir (2009) from a corpus of texts from 143 texts from five different exams including FCE. Finally, the remainder came from a study commissioned for the publication of the Khalifa and Weir (2009) volume, with the data comprising 30 reading texts from five different exams including FCE. The total text length is thought to be indicative of cognitive demand, whereas variables from 5 onwards

relate most to lexical access (Khalifa & Weir, 2009). The reminder of the variables are most likely to relate indirectly to the concept of text complexity which combines a variety of causes. For example, a larger number in mean words per sentence could indicate a more difficult text. The cause of the difficulty may be increased demand at the syntactic parsing stage of processing, but may also indicate a more lexically dense text (Weir, 2013).

In order to prepare the test tasks for automated analysis, the text for each was edited using Word (Microsoft Corp., 2010b) as set out in Table 17.

Table 17 Editing of test task texts

Part	Editing
1	The keys were added as titles to the text at the point specified by the task
2	No editing
3	The keys were added to fill the gaps in the text at the point specified by the task
4	The four texts were grouped together as paragraphs of one text

To produce the indices for variables 1 to 4, Coh-Metrix 3.0 (McNamara et al., 2012) was used. For Coh-metrix analyses, it is necessary to set the genre using a drop down box on the website. These settings calibrate the analysis for those indices requiring reference texts. For genre, selection can be made from 'science', 'narrative' and 'informational'. It is recommended that selection be made according to the term which matches the text most closely. As the texts were not scientific and had features of a narrative structure, such as characters and time references, 'Narrative' was chosen. The texts were then analysed processed by the system and the results downloaded in the form of an Excel Workbook (Microsoft Corp., 2010a).

For variables from 5 onwards, each the texts, prepared as described in Table 17, was processed by VocabProfile (Cobb, 2013). Both BNC 20 and AWL analyses were used. The results of this analysis were summarised to match information available in Khalifa and Weir (2009), which was used as a reference set.

To compare the test overall with the figures provided by Khalifa and Weir (2009), further statistics were created from those for each test part. For the overall number of words, tokens and types, the figures for each part were summed. The mean was taken for the two Flesch readability statistics and lexical density. For all others, figures were proportions, and these calculations were done based on the sum of the totals for each part. For example, the combined figure for K1 words was the sum of the total number of words in at that level in each text divided by the sum total number of words in each text. The results can be found in 4.2.3, with the unabridged output from Coh-Metrix and VocabProfile available in Appendix 9: descriptive statistics for test materials.

3.5 Further analysis of the test materials

3.5.1 Expert judgement of relevant text for each option

In order to obtain the segment(s) of text of particular importance to each item option (Embretson & Wetzel, 1987), three experts were asked to analyse the test tasks and posit the segments of text they felt might lead a candidate to select each option. Henceforward, this text will be referred to as the *relevant text* for an item. Each expert had worked in the field of language testing for more than five years, and, among other responsibilities, was required to appraise the effectiveness of items; two of the three experts were also item writers. Editing and/or writing items was considered important to match the judgement task to judge expertise (see 2.9.2). This was because the aim of the exercise was to recover the text from the reading passage which was crucial in responding to items. Item editors and writers routinely consider this when working on items.

The procedure was conducted as follows. Experts were provided with the test materials and a form in which to record their judgements (see Appendix 10: instructions for selection of relevant text). They were asked to complete this process using a computer by copying and pasting the relevant text for each option into the appropriate cell of the form.

Where the segments of text selected by the experts included partial sentences, the highlighted sections were expanded to include the entire sentence, so that, in

subsequent analyses, only whole sentences were considered. This was done to provide a consistent rule to use when collating judgements and because some of the indices to be used (e.g. the number of sentences) presuppose whole sentences. This modification was also justified on the grounds that experts were thought likely to select the same core text as each other, regardless of the exact boundaries they drew. Furthermore, given the numbers of candidate responses, and the variation of response processes likely, it was decided that attempting a higher level of precision when circumscribing the relevant text would not produce more accurate results. Instead, a greater, rather than lesser range of text was favoured as a general principal. This was because, concerning the estimation of indices, including text which is relevant would be at least as important, if not more so than excluding text which was irrelevant. For example, indices, such as those which represent the density of a feature (e.g. the number of propositions per word) would be less accurate if irrelevant text were included, but the effect of adding more words would be relatively trivial because the index is, in effect, averaged over all the words, including those actually of interest. Other types of index, such as those measuring lexical frequency would at least include all the information of importance, even if unimportant information were included as well.

In order to validate the work of the expert judges, an agreement index was calculated. This was the number of actual agreements between experts divided by the number of opportunities for agreement. A single opportunity for agreement was defined as a sentence selected by one expert for a particular item option. An actual agreement was counted if another expert selected the same sentence for the same item option. If no sentences were selected by any expert for an option, this was considered to be a single agreement. The result was a figure expressing the proportion of agreement among the raters. An overall figure of more than 0.5 was considered sufficient to continue as this would indicate more agreement than disagreement. Results are reported in 4.3.1.1.

3.5.2 Determination of relevant text for subsequent analysis

For the reasons mentioned in 3.5.1, after the selection of relevant text by the experts, where incomplete sentences were selected, they were re-interpreted as indicating complete sentences (see 3.5.1). The separate judgements of experts were then combined as follows. For clarity, illustrated cases are provided in

Table 18 and noted in the description.

Case A: where judgements were in close agreement (i.e. most of the sentences were the same, although some experts may have nominated additional sentences connected to those agreed upon), both the agreed text and the additional sentences were accepted, following the principle of greater range, discussed in 3.5.1.

Cases B, C: where experts selected quite different text, the majority decision was followed, with the greater range favoured, as before, when comparing the selections of the majority.

Case D: in cases where only one expert selected any text at all, this text was adopted.

Case E: if more than one expert had selected text but there was no agreement, what seemed to be the most reasonable text to the researcher (me) would be selected.

Case F: if nothing was selected by any expert, no text would be adopted.

Table 18 Illustration of rules for combining expert judgements on relevant text

Case	Expert 1	Expert 2	Expert 3	Outcome
A	A	A	A	A selected
B	A	A	B	A selected
C	A	A		A selected
D	A			A selected
E	A	B	C	Researcher selects from A, B or C
F				Nothing selected

3.6 Construction of task process indicators

3.6.1 Attribute indicators

3.6.1.1 Indicator target

Indicators were to relate to one of the following:

- the key of each item
- the distractors of each item
- both the key and distractors of each item

Since indicators were to be derived from one of several processes, most requiring specific input text, relevant text was selected according to the target of the indicator listed above. In a few cases, such as with the propositionalisation of the text, the whole text was the input to the analysis process, and the output divided according to the target of the each item's relevant text.

3.6.1.2 Analysis

The indicators constructed were derived using several different sources. These methods are listed in Table 19, together with the number of times they were used. A more detailed description follows in 3.6.1.3, 3.6.1.4, 3.6.1.5 and 3.6.1.6, with a detailed summary in Table 20, Table 21, Table 22, Table 23, Table 24, Table 25, Table 26, Table 27, Table 28, Table 29 and Table 30.

Table 19 Frequency of use of methods to create indicators

Method	Number of indicators created	Indicators[10]
Coh-metrix (McNamara et al., 2012) indices	36	OP, READ: number of syllables, content word frequency CELEX, all word frequency CELEX, content word frequency log CELEX, type-token ratio, hypernymy, polysemy, lexical density, concreteness, modifiers per noun phrase, left embeddedness, negation, passive voice, connectives, stem overlap, proposition density READ: causality, intentionality, temporality, sentences
VocabProfile (Cobb, 2013) indices	4	OP, READ: maximum frequency BNC, maximum frequency AWL
CPIDR 5.1 (Brown et al., 2012) proposition estimates	4	OP, READ: propositions, proposition density
LSA (Laham, 1998) indices	5	SEARCH: LSA by option by term RD: LSA key by term, LSA distractor by term, LSA key by document, LSA distractor by document,
Expert judgement	4	OP, READ: holistic negation, fronted sentences
The application of rules (e.g. counting the total number of words between relevant test for a particular item, as in 58 (X058.RD.disperse), **Table 29 and Table 30)**	4	SEARCH: search order, demarcatedness RD: relevant text dispersal, practice effect,
Total	57[11]	

[10] Indicators, such as proposition density, appearing in more than against more than one method do so because they were created using information from diverse categories.

[11] There were 55 initial indicators but two required input from both Coh-Metrix and CPIDR hence a total of 57 in this table.

3.6.1.3 Coh-metrix

For Coh-metrix analyses, it is necessary to set the genre using a drop down box on the website. 'Narrative' was chosen for the reasons given in 0.

3.6.1.4 VocabProfile indices

The indices provided by VocabProfile (Cobb, 2013) represent the frequency level of the word using two corpora: the British National Corpus (BNC) and the Academic Word List (AWL). Frequency levels provided for the BNC consist of the most common 1,000 words form the first level, the next most common 1,000 words form the next level, and so on. For example, a word designated with *1* means that the word is among the thousand most frequent words in that corpus; a *2* means it falls within the second most frequent 1,000 words. Tiers range from 1 to 25, with 26 indicating off-list words[12]. The AWL index has four levels. The first two denote the first 1,000 and 2,000 most frequent words from the BNC after AWL words are excluded. The third category comprises words on the AWL and the fourth off-list words. If a word appears in both the first of second 1,000 words and the AWL, it is marked as belonging to the AWL. For this study, the highest value found in the relevant text was taken (Graesser et al., 2011), thus assuming that unlisted words were more difficult than listed words. More low frequency words therefore result in a higher index.

3.6.1.5 CPIDR 5.1

This involves the automatic propositionalisation of text based on parts of speech parsing. Propositions are identified in relation to parts of the sentence, and so can be summed for each sentence. Summing for each sentence was completed using Excel (Microsoft Corp., 2010a).

3.6.1.6 LSA indices

LSA can link texts by comparing each term in one text to each term in another text. This is closely related to the processes involved in searching for relevant text: locating words, semantic approximations or topics (Khalifa & Weir, 2009). During response decision, selecting the correct responses is thought to be based on

[12] See Weir (2013) for justification for treating off list as difficult words.

comparing the various plausible options, and other information related to the item, using the situational model which has been generated through careful reading. For this reason, document to document matching was selected, as it attempts to relate the summed meaning of the document's terms to that of the other document. However, it is possible that candidates also use the textbase to make their response decisions at this stage, so a term to term indicator was also included. There is an additional requirement to select *topic space* for each analysis. Topic space refers to the corpus of texts which was used to establish that particular set of semantic relationships between words, as they are expected to vary with contextual factors (Dennis, 2011). The default topic space is 'General Reading up to 1st year college', other general reading topic spaces refer to 3rd, 6th, 9th and 12th grade. In all cases, the topic space selected was 'General Reading up to 1st year college'. This is because the texts in the current study were clearly general reading, and up to 1st year college was chosen as the task texts were not graded for any particular level.

3.6.1.7 Expert judgement indices

Two experts were invited to analyse the texts. In both cases, the experts had completed Masters Degrees in applied linguistics, had been teachers of English for more than five years, had spent more than five years in language testing and reported themselves to be familiar with textual analysis. In respect of the match between judge and the task they are asked to complete (2.9.2), expertise in textual analysis and the ability to understand the definition of the attributes required, was considered important. In both cases, the Masters degrees completed and the experience in teaching ensured that the judges were suitable. They were presented with the texts and item stems for each task and asked to determine the following independently of each other:

- the incidence of negations of all types (not only grammatical)
- the incidence of fronted structures

The expert judgment in this case was validated by computing the number of agreements for each attribute as a proportion of all judgements. This was done for each attribute at task and overall level. The extent of agreement is reported in 0. In the case of complete agreement being achieved, all judgements would be accepted. In the event of any disagreements, experts were asked to discuss differences and make a joint decision on all judgements.

3.6.1.8 Combining

In all cases, an item-level indicator was required for the main analysis. For some indicators, this was the output of the analysis stage (such as with the Coh-metrix indicators relating to the key alone e.g. the LSA match between key and relevant text). In other cases, word, proposition, sentence or option level information had to be combined. In most cases, combining was done either by a raw count of the phenomena in question over the relevant text in question (such as with word-level indices), or by taking the mean of the indices to be combined (such as with most option-level indices), as this ensured the resulting indicator was representative of all constituent indices. Full details of how indices were constructed is presented for each indicator in Table 20, Table 21, Table 22, Table 23, Table 24, Table 25, Table 26, Table 27 and Table 28. A total of 55 indicators resulted from these steps.

Table 20 Item attribute indicators OP – basic characteristics

	Name	Subcomponent	Gloss	Expected impact on item difficulty
1	X001.OP.syll	Word recognition	Number of syllables	More difficult (-ve coefficient)
2	X002.OP.BNC	Lexical access	Maximum frequency BNC	More difficult (-ve coefficient)
3	X003.OP.AWL		Maximum frequency AWL	More difficult (-ve coefficient)
4	X006.OP.CELEX.cont.f		Content word frequency CELEX	Easier (+ve coefficient)
5	X007.OP.CELEX.all.f.log		All word frequency CELEX	Easier (+ve coefficient)
6	X008.OP.CELEX.cont.log		Content word frequency log CELEX	Easier (+ve coefficient)
7	X010.OP.hypernymy		Hypernymy	More difficult (-ve coefficient)
8	X011.OP.polysemy		Polysemy	More difficult (-ve coefficient)
9	X012.OP.lex.density		Lexical density	More difficult (-ve coefficient)
10	X013.OP.concrete		Concreteness	Easier (+ve coefficient)
11	X014.OP.mod.noun	Syntactic parsing	Modifiers per noun phrase	More difficult (-ve coefficient)
12	X015.OP.left.emb		Left embeddedness	More difficult (-ve coefficient)
13	X016.OP.neg	Establishing propositional meaning	Negation	More difficult (-ve coefficient)
14	X017.OP.hol.neg		Holistic negation	More difficult (-ve coefficient)
15	X018.OP.fronted		Fronted sentences	More difficult (-ve coefficient)
16	X019.OP.passive		Passive voice	More difficult (-ve coefficient)
17	X022.OP.props		Propositions	More difficult (-ve coefficient)
18	X000.OP.prop.dens		Proposition density	More difficult (-ve coefficient)

Table 21 Item attribute indicators OP – processing I

	Name	Input	Initial analysis	Further manipulation
1	X001.OP.syll	Stem and option text	Coh-metrix (8 DESWLsy 'Word length, number of syllables, mean')	None
2	X002.OP.BNC	Stem and option text	VocabProfile (identification of frequency tier in BNC for each word)	Max freq. tier
3	X003.OP.AWL	Stem and option text	VocabProfile (identification of frequency tier in AWL for each word)	Max freq. tier
4	X006.OP.CELEX.cont.f	Stem and option text	Coh-metrix (94 WRDFRQc 'CELEX word frequency for content words, mean)	None
5	X007.OP.CELEX.all.f.log	Stem and option text	Coh-metrix (95 WRDFRQa 'CELEX log frequency for all words, mean)	None
6	X008.OP.CELEX.cont.log	Stem and option text	Coh-metrix (96 WRDFRQmc 'CELEX log minimum frequency for content words, mean)	None
7	X010.OP.hypernymy	Stem and option text	Coh-metrix (105 WRDHYPnv 'Hypernymy for nouns and verbs, mean')	None
8	X011.OP.polysemy	Stem and option text	Coh-metrix (102 WRDPOLc 'Polysemy for content words, mean')	None
9	X012.OP.lex.density	Stem and option text	Coh-metrix (84 (WRDNOUN 'noun incidence') + 85 (WRDVERB 'verb incidence') + 86 (WRDADJ 'adjective incidence') + 87 (WRDADV 'adverb incidence') divided by 3 (DESWC 'word count))	None

Table 22 Item attribute indicators OP – processing II

	Name	Input	Initial analysis	Further manipulation
10	X013.OP.concrete	Stem and option text	Coh-metrix (99 WRDCNCc 'Concreteness for content words, mean')	None
11	X014.OP.mod.noun	Stem and option text	Coh-metrix (70 SYNNP 'Number of modifiers per noun phrase, mean')	None
12	X015.OP.left.emb	Stem and option text	Coh-metrix (69 SYNLE 'Left embeddedness, words before main verb, mean')	None
13	X016.OP.neg	Stem and option text	Coh-metrix (81 DRNEG 'Negation density, incidence')	None
14	X017.OP.hol.neg	Stem and option text	Expert judgement (identification of number of negations per sentence)	Count
15	X018.OP.fronted	Stem and option text	Expert judgement (identification of sentences with fronted structures)	Count
16	X019.OP.passive	Stem and option text	Coh-metrix (80 DRPVAL 'Agentless passive voice density, incidence')	None
17	X022.OP.props	Stem and option text	CPIDR 5.1	Count
18	X000.OP.prop.dens	Stem and option text	X022.OP.props divided by Coh-metrix 3 (DESWC 'word count')	None

Table 23 Item attribute indicators SEARCH – basic characteristics

	Name	Subcomponent	Gloss	Expected impact on item difficulty
19	X051.SEARCH.order	Order	Search order	Easier (+ve coefficient)
20	X052.SEARCH.demarc	Demarcation	Demarcatedness	Easier (+ve coefficient)
21	X053.SEARCH.LSA.term	LSA match	LSA by term	Easier (+ve coefficient)

Table 24 Item attribute indicators SEARCH – processing

	Name	Input	Initial analysis	Further manipulation
19	X051.SEARCH.order	Item text + task text	Application of rule (identification of whether relevant text for items followed order of items)	None
20	X052.SEARCH.demarc	Task text	Application of rule (identification of whether the relevant text was demarcated in some way)	None
21	X053.SEARCH.LSA.term	Stem and option + relevant text	LSA (term to term comparison)	None

Table 25 Item attribute indicators READ – basic characteristics I

	Name	Subcomponent	Gloss	Expected impact on item difficulty
22	X026.READ.syll	Word recognition	Number of syllables	More difficult (-ve coefficient)
23	X027BNC	Lexical access	Maximum frequency BNC	More difficult (-ve coefficient)
24	X028AWL		Maximum frequency AWL	More difficult (-ve coefficient)
25	X031.READ.CELEX.cont.f		Content word frequency CELEX	Easier (+ve coefficient)
26	X032.READ.CELEX.all.f.log		All word frequency CELEX	Easier (+ve coefficient)
27	X033.READ.CELEX.cont.log		Content word frequency log CELEX	Easier (+ve coefficient)
28	X034.READ.type.tok	Establishing a coherent textbase	Type-token ratio	More difficult (-ve coefficient)
29	X035.READ.hypernymy	Lexical access	Hypernymy	More difficult (-ve coefficient)
30	X036.READ.polysemy		Polysemy	More difficult (-ve coefficient)
31	X037.READ.lex.density		Lexical density	More difficult (-ve coefficient)
32	X038.READ.concrete		Concreteness	Easier (+ve coefficient)
33	X039.READ.mod.noun	Syntactic parsing	Modifiers per noun phrase	More difficult (-ve coefficient)
34	X040.READ.left.emb		Left embeddedness	More difficult (-ve coefficient)

Table 26 Item attribute indicators READ – basic characteristics II

	Name	Subcomponent	Gloss	Expected impact on item difficulty
35	X041.READ.neg	Establishing propositional meaning	Negation	More difficult (-ve coefficient)
36	X042.READ.hol.neg		Holistic negation	More difficult (-ve coefficient)
37	X043.READ.fronted		Fronted sentences	More difficult (-ve coefficient)
38	X044.READ.passive		Passive voice	More difficult (-ve coefficient)
39	X045.READ.connect	Establishing a coherent textbase	Connectives	Easier (+ve coefficient)
40	X046.READ.stem.o		Stem overlap	Easier (+ve coefficient)
41	X047.READ.props	Establishing propositional meaning	Propositions	More difficult (-ve coefficient)
42	X000.READ.prop.dens		Proposition density	More difficult (-ve coefficient)
43	X048.READ.causal	Building a situational model	Causality	Easier (+ve coefficient)
44	X049.READ.intent		Intentionality	Easier (+ve coefficient)
45	X050.READ.temp		Temporality	Easier (+ve coefficient)
46	X000.READ.sentence	Establishing a coherent textbase	Sentences	More difficult (-ve coefficient)

Table 27 Item attribute indicators READ – processing I

	Name	Input	Initial analysis	Further manipulation
22	X026.READ.syll	Relevant text	Coh-metrix (8 DESWLsy 'Word length, number of syllables, mean')	None
23	X027BNC	Relevant text	VocabProfile (identification of frequency tier in BNC for each word)	Max freq. tier
24	X028AWL	Relevant text	VocabProfile (identification of frequency tier in AWL for each word)	Max freq. tier
25	X031.READ.CELEX.cont.f	Relevant text	Coh-metrix (94 WRDFRQc 'CELEX word fequency for content words, mean)	None
26	X032.READ.CELEX.all.f.log	Relevant text	Coh-metrix (95 WRDFRQa 'CELEX log requency for all words, mean)	None
27	X033.READ.CELEX.cont.log	Relevant text	Coh-metrix (96 WRDFRQmc 'CELEX log minimum fequency for content words, mean)	None
28	X034.READ.type.tok	Relevant text	Coh-metrix (48 LDTTRc 'Lexical diversity, type-token ratio, content word lemmas')	None
29	X035.READ.hypernymy	Relevant text	Coh-metrix (105 WRDHYPnv 'Hypernymy for nouns and verbs, mean')	None
30	X036.READ.polysemy	Relevant text	Coh-metrix (102 WRDPOLc 'Polysemy for content words, mean')	None
31	X037.READ.lex.density	Relevant text	Coh-metrix (84 (WRDNOUN 'noun incidence') + 85 (WRDVERB 'verb incidence') + 86 (WRDADJ 'adjective inceidence') + 87 (WRDADV 'adverb incidence') divided by 3 (DESWC 'word count))	None
32	X038.READ.concrete	Relevant text	Coh-metrix (99 WRDCNCc 'Concreteness for content words, mean')	None
33	X039.READ.mod.noun	Relevant text	Coh-metrix (70 SYNNP 'Number of modifiers per noun phrase, mean')	None
34	X040.READ.left.emb	Relevant text	Coh-metrix (69 SYNLE 'Left embeddedness, words before main verb, mean')	None

Table 28 Item attribute indicators READ – processing II

	Name	Input	Initial analysis	Further manipulation
35	X041.READ.neg	Relevant text	Coh-metrix (81 DRNEG 'Negation density, incidence')	None
36	X042.READ.hol.neg	Relevant text	Expert judgement (identification of number of negations per sentence)	Count
37	X043.READ.fronted	Relevant text	Expert judgement (identification of sentences with fronted structures)	Count
38	X044.READ.passive	Relevant text	Coh-metrix (80 DRPVAL 'Agentless passive voice density, incidence')	None
39	X045.READ.connect	Relevant text	Coh-metrix (52 CNCAll 'All connectives incidence')	None
40	X046.READ.stem.o	Relevant text	Coh-metrix (30 CRFSO1 'Stem overlap, adjacent sentences, binary, mean')	None
41	X047.READ.props	Relevant text	CPIDR 5.1	Count
42	X000.READ.prop.dens	Relevant text	X047.READ.props divided by Coh-metrix 3 (DESWC 'word count)	None
43	X048.READ.causal	Relevant text	Coh-metrix (64 SMCAUSr 'Ratio of casual particles to causal verbs')	None
44	X049.READ.intent	Relevant text	Coh-metrix (65 SMINTEr 'Ratio of intentional particles to intentional verbs')	None
45	X050.READ.temp	Relevant text	Coh-metrix (68 'temporal cohesion')	None
46	X000.READ.sentence	Relevant text	Coh-metrix (02 ‘Number of sentences')	None

Table 29 Item attribute indicators RD – basic characteristics

	Name	Subcomponent	Gloss	Expected impact on item difficulty
47	X054.RD.LSA.term.KEY	LSA match	LSA key by term	Easier (+ve coefficient)
48	X055.RD.LSA.term.DIST		LSA distractor by term	More difficult (-ve coefficient)
49	X056.RD.LSA.doc.KEY		LSA key by document	Easier (+ve coefficient)
50	X057.RD.LSA.doc.DIST		LSA distractor by document	More difficult (-ve coefficient)
51	X058.RD.disperse	Dispersal	Relevant text dispersal	More difficult (-ve coefficient)
52	X059.RD.pract	Practice effect	Practice effect	Easier (+ve coefficient)

Table 30 Item attribute indicators RD – processing

	Name	Initial analysis	Further manipulation
47	X054.RD.LSA.term.KEY	LSA (term to term comparison)	None
48	X055.RD.LSA.term.DIST	LSA (term to term comparison)	None
49	X056.RD.LSA.doc.KEY	LSA (doc to doc comparison)	None
50	X057.RD.LSA.doc.DIST	LSA (doc to doc comparison)	None
51	X058.RD.disperse	Application of rule (total number of words constituting the plausible text for each item)	None
52	X059.RD.pract	Application of rule (frequency of sentence usage prior to reuse in connection to plausible text)	None

3.7 Construction of other indicators and matrices necessary for the analysis

3.7.1 Incidence matrix

The item by indicator matrix containing the attribute indicators is termed the *incidence matrix*. The first indicator in each row of the matrix was a factor, or grouping variable, to identify each item. In order to specify items nested in tasks (belonging to them exclusively), a task factor was created. It consisted of a number between 1 and 4 identifying the task (Part) each item belonged to.

The incidence matrix finished by containing a combination of factors and continuous variables as indicators. GLMM can handle a mixture of both types of indicator without a problem. An understanding of both is required, however, when interpreting the results. Indicators based on continuous variables represent within item variance, as the variable is represented in each item (the cases) to a varying extent.

A factor is a grouping variable with at least two levels. The cases, in the current research, items, belong to one of these levels exclusively. Factors may assume the levels are ordered in some way but ordering was not specified in the current analysis because it was unnecessary. If an order existed in the data, it would be apparent in the fixed effect coefficients for each level. One of the levels may also represent absence of the influence of an attribute. This may mean that the indicator represents between item variance only (with two levels: absence and presence) or a combination of between item variance and within item variance (with at least three levels: absence, presence of type A, presence of type B).

One important difference between continuous and factor-based indicators is the amount of information each contains. Unlike for factors, continuous variables assume that intervening values between the values specified for variables exist and indicate relative positions on an ordered scale. Factors are far cruder measurements, only able to group similar cases and to order these groups. This can be problematic because factors cannot account for nuanced differences between attributes and, with a small number of cases, small differences may be an important distinguishing feature.

For both types of indicator, there is the further problem of influential data points, which will have more impact in an incidence matrix with a small number of cases. This is not a problem when exclusively considering a single data set, as the values for coefficients summarise the overall impact of indicators on the probability of success. If the value for one attribute were particularly influential, it would affect the results of any candidate taking the test, so it would be fair to include it. Generalisability to other forms of the same test may be affected, however, as the

values of some indicators may be less influential in the broader context of several test forms. This issue will be discussed in the when reviewing the generalisability of findings in 5.2.

3.8 Main analysis

3.8.1 Collation of the data

The main analysis was carried out using the *lme4* (Bates et al., 2014) package for R (R Core Team, 2014). lme4 (Bates et al., 2014) requires data in long format, whereby each row contains all indicator values relevant to a particular data point (a score) and that data point, as opposed to the wide data format of the matrices. In the wide matrix which contains 10,000 candidates and 35 items, its 350,000 scores (one per cell) would each be represented on its own row in long format. The melt() function of another R package, *reshape* (Wickham, 2013), was used, as suggested by De Boeck et al. (2011), to transform the score data. The result was a matrix with a row for each response which also contained the relevant candidate and item Identification numbers (person and item indicators). The merge() function was then used to append the information in the incidence matrix using the 'item' identification number to link the data.

3.8.2 The development of a model for statistical analysis

3.8.2.1 Fitting a Rasch model to the data

Fischer (1995) advised fitting a Rasch model to the data and testing assumptions of unidimensionality and LD before fitting a LLTM. This was two reasons. First, because both models are similar but the Rasch model was easier to implement, and far more likely to fit the data (De Boeck & Wilson, 2004a), if the Rasch model violated assumptions, the LLTM could be assumed to do so also. In this case, as discussed in 2.9.3.3.4, a solution would be to add dimensions to the model to create a random weights Rasch model, just as would be done to change a LLTM into a RWLLTM. The second reason for fitting a Rasch model is that it provides a means of comparison for the LLTM or RWLLTM after all models have been estimated. This approach is quite common when using LLTM, as, if the attributes specified in the LLTM are facets of the item difficulty parameter in the Rasch model, the difference in the amount of variance explained by the two models is

the capacity of the attributes to explain item variance (De Boeck & Wilson, 2004a; Wilson & Moore, 2012).

A Unidimensional Rasch model was specified, with the items as fixed effects (De Boeck et al., 2011). The data were investigated for unidimensionality and local item independence by analysis of the residuals (the difference between observed and values expected by the model) produced by the analysis. Residuals, rather than the score data are recommended because it is possible for so-called difficulty factors to be detected with methods where item difficulty has not already been accounted for and it was secondary dimensions which were of interest (Smith, 1996), rather than that accounted for by the Rasch model. As suggested by DeMars (2010), scree plots, which show the amount of variance due to each potential dimension in descending order, were constructed using SPSS (IBM Corp., 2013) and analysed. The plots are so-named as they resemble the profile of the join between a mountain slope and the surface of the earth, where the slope of the mountain is lessened by scree fall. The number of meaningful dimensions in the data is interpreted as the number of points plotted before the biggest drop (DeMars, 2010). In the case where the data contained a single dimension, and this was already accounted for by the Rasch model, the residuals would show an entirely flat profile without a significant drop between dimensions. If a profile with a steep drop were found, dimensions, in the form of random weights, would be added and the model tested again. The result might not be a perfectly flat scree plot but, in accordance with the discussion in 2.9.3.3.2, it would be important that the variance explained by these dimensions be reduced and equalised so that any large drop in the chart would disappear.

There was an expectation that, if dimensions were found through examination of the residuals, they would correspond to the four test tasks. In this case, it was thought that the addition of dimensions would account for most LD, if found (see 2.9.3.3.2). For this reasons, dimensionality corrections were dealt with first and attention turned to correcting LD after that. Nevertheless, local item independence was investigated for all models, as this would serve as a useful

reference if LD were found after corrections for dimensionality. In order to detect LD, the Q_3 index was calculated (Yen, 1984) and then squared to produce Q_3^2, as suggested by De Ayala (2009). Q_3 is simply a Pearson correlation of the residuals from an analysis. Residuals which correlate highly (either negatively, or positively) indicate a dependence between items which is not modelled. The index is squared to aide interpretation: the result of a squared correlation is R^2, the proportion of shared variance explained by the index. The expectation for most item pairs would be around zero. If any values were above 0.15, a correction would be applied to the highest pairing and the model rerun to determine whether any further correction was required. The remedy for dependency between pairs of items exhibiting LD was that described by De Boeck et al. (2011). Fixed effects were added to the model for problematic pairs of items, whereby a 1 is recorded for dependency between the items where a candidate had the correct response, otherwise a 0 was used. An acceptable outcome was where no item pair stood out as higher than a corrected pairing. This was done, rather than setting an arbitrary cut off point, as the aim was to equalise violations and thus reduce their impact (see 2.9.3.3.2).

Throughout the process of fitting different Rasch models, comparative fit was also examined. This was important to establish that a modified model fitted significantly better than the unmodified model. All models with additional parameters are expected to fit the data better, even if the additional parameters are randomly specified. However, if the magnitude of the improvement is greater than that which would be expected to occur by chance indicates a better model (Gelman & Hill, 2006). The Likelihood Ratio Test (De Boeck et al., 2011; DeMars, 2010; Gelman & Hill, 2006) was used to determine this where applicable. The test is suitable for nested models, such that the model with fewer parameters only contains parameters which are also found in the larger mode. It can therefore be viewed as a restricted version of the larger model and the LRT as a test where the null hypothesis that the additional parameters fit better due to chance. The initial unidimensionality Rasch model was also tested against a so-called *empty model*, which contained no fixed effects but the same random effects.

Two further assumptions relating to the testing of these models were also discussed in 2.9.3.3.4. The first was normality of the distribution of person estimates. This was tested only in for models for which the first two assumptions were found to hold. It was tested by examining the mean, median, skew and kurtosis of the values for each dimension of a model. A mean and median which are close to each other, and between 2 and -2 (Bachman, 2004) for skew and kurtosis were taken to indicate approximate normality. The second assumption was the absence of colinearity between fixed effects. Since the Rasch models specified in this study contain items as fixed effects, the absence of colinearity assumption is already dealt with when the local item independence assumption is addressed. For this reason, it was not investigated further for Rasch models.

Results of the model specification phase can be found in 4.4.1.

3.8.2.2 Selection of indicators

Having found a suitable Rasch model which met assumptions, it was used as the basis for subsequent LLTM models. In other words, any dimensions or fixed effects added to account for violations of unidimensionality or local item independence were retained, and the fixed effects for items replaced with those for attributes. It was then necessary to test indicators to determine their qualities and which should be retained for the final model. Thus, two phases of model testing were adopted:

1. testing indicators
2. testing and comparing composite models

Indicators were tested for two reasons: i) to provide information about individual indicators, ii) to provide evidence for the selection of indicators for later composite models. To test indicators, a model for each indictor was specified in order to avoid the problem of colinearity, or shared variance, between pairs of indicators which can distort estimates of individual indicators, although this does not affect indices and statistics concerning the model as a whole (De Boeck et al., 2011).

There are two further advantages of specifying one model for each indicator. The first related to the way in which coefficients for factor indicators are reported by the software (Bates et al., 2014) and the way in which this affects other coefficients. Estimates for factor indicators are expressed for each level except one which is used as a reference level. Estimates for all indicators, factor and numeric, are also expressed in relation to this reference level, which makes them slightly more difficult to interpret. If there is more than one factor in a model, several reference levels (one for each factor) must be considered when interpreting the indicators and interpretation quickly becomes complicated. The intercept of the first factor in a model may be supressed, and an estimate for each level of that factor generated but this is not possible with the other factors.

The second advantage of specifying one model for each indicator is to avoid issues of identification. Identification concerns technical requirements of mathematical models which help to ensure that the parameters estimated when the model is fitted to the data are unique. The alternative would be an output of one set of estimates which could just as easily have been any from a wide range of other values (Kenny & Milan, 2012). Clearly, estimates which are not uniquely possible given a model and data set are of little use as research findings. Among the technical requirements for identification of LLTM are that the number of indicators should be fewer than the number of cases (items) and that both the score matrix and the indicator matrix must be of full column rank. Full columns rank means that it is impossible to recreate any of the columns of the matrix by multiplying a combination of the other columns in the same matrix. The chances of violating this requirement increase and the number of columns in a matrix increases because the possible combinations of columns which can be multiplied can be increased.

For each indicator, a LLTM with the modification determined in 3.8.2.1 was specified. Indicators were appraised on the basis of the fixed effect coefficient and its statistical significance, as described in 3.8.2.2.1 for indicators based on continuous variables and in 3.8.2.2.2 for those based on factors, or categorical variables.

3.8.2.2.1 Continuous indicators

3.8.2.2.1.1 Fixed effect significance

Each coefficient estimated by the model was also tested for significance using a z-test, produced automatically by the software. A statistically significant coefficient is one which is very unlikely to be due to chance, and therefore highly likely to represent the impact of the indicator. Coefficients which were interpretable and significant were retained. The significance level chosen for testing was <0.1, which is somewhat more liberal than typically adopted alternatives, such as 0.05. This was done because all the indicators selected for testing were potentially relevant. An indicator with relatively mild significance is unlikely to influence the overall model very much, as it would typically have a very weak coefficient.

3.8.2.2.1.2 Fixed effect coefficient

The substantive impact of an indicator is estimated by the coefficient calculated during the analysis. The coefficient represents the influence of the indicator over the probability of success on all items in the analysis. A positive coefficient is estimated when the indicator makes items easier, a negative coefficient when the indicator contributes to difficulty. In each case, the probability is expressed as a variation from a probability of success of 0.5 on the logistic scale (De Boeck et al., 2011)

Those indicators found to influence item difficulty as predicted by theory (see 2.7, summarised in Table 20, Table 23, Table 25, Table 26, Table 29) were examined for a value which was consistent with the theory it was based upon. Indicators which did not behave as predicted by theory were considered uninterpretable for the purposes of the current study and omitted. This is not an uncommon practice, Buck et al. (1997) and Aryadoust and Goh (2014), for example, rejected variables on substantive grounds. The current study does not attempt to re-write prevailing cognitive theory, but, like Buck et al. (1997) and Aryadoust and Goh (2014), use its support to explain the data. For this reason, the approach must be theory driven and the dropping of unexplainable results is considered essential. Such an approach, is common to many analytical approaches, where parsimony in interpretation is considered essential (Chou & Huh, 2012).

A further reason for not including indicators in the final model was also based on substantive grounds. Several indicators were measuring essentially the same attribute, albeit in somewhat different ways. Most obvious were those indicators measuring lexical frequency, of which there were five OP components and five corresponding indicators for the READ component (see 3.6.1.8). Each was intended to be measuring item demands on the process of lexical access through word frequency, but any differences between results could not be explained in terms of slight differences in the cognitive process measured. Instead, reasons such as the appropriacy of the corpus upon which the indicator was based, or the process by which it was derived were more obvious. A consequence of retaining more than one indicator where the differences could not be explained on a theoretical basis would be a better fitting model with no theoretical gain. For this reason it was avoided and only the indicator accounting for the most variance (having the largest absolute coefficient) in each case was retained.

3.8.2.2.2 Factor-based indicators

The examination of factor-based indicators is the same as for continuous indicators with the addition of two further steps. As factors divide items into two or more levels, a coefficient for each level is produced. The magnitude of these coefficients is expected to follow an expected order so that increasing quantities of the variable correspond to greater, or lesser difficulty depending on theory. A further requirement was that the coefficients for each level were statistically separable. In other words, a confidence interval was created for each coefficient by adding and subtracting the value for error multiplied by two. The value for error is multiplied by two, as this approximates a 95% confidence interval. If the confidence interval overlapped for any two coefficients, the coefficients were not considered statistically separable.

It was also possible, that after examination of the results of factor based indicators, there could be some evidence that collapsing the categories might aid interpretation. This might be in cases, for example, where coefficients were for adjacent levels were very close but did not form a monotonic pattern overall. By combining such categories, and testing the new indicators, it was possible that the

result would be interpretable according to the method presented here. Such an approach is reasonable where the indicator is not simply presence or absence of an attribute. This is because the dividing line which separates one category from another is somewhat arbitrary due mainly to the crudeness of the scale (3.7). Specific reasons for the creation of new categories are given with the results of their parent categories in 0, where the results can also be found.

3.8.2.2.3 Presentation of results

Results for the testing of all fixed effects are available in 0. The coefficient is expressed in terms of the contribution of the fixed effect on the log-odds of success on test items. This is also expressed as the probability of success for convenience. If this figure is below 0.5, the fixed effect has a negative impact on success (items are harder). In order to express this figure relative to 0, 0.5 is subtracted from the probability and the result expressed as a percentage. This figure is labelled 'influence'. For factor-based indicators, the level where least impact is expected was set as the reference level and given a value of 0. Influence for all other levels is, therefore, expressed relative to this, such that greater difficulty is represented by negative numbers and reduced difficulty by positive numbers.

3.8.3 Final model

A final model was specified to gauge the effectiveness of the selected indicators in explaining overall variance in the current study. To specify this model, indicators were selected on the basis of their performance in the previous stage, according to the significance and substantive impact of the fixed effect coefficients (3.8.2).

3.8.3.1 Appraising the final model

3.8.3.1.1 Testing model assumptions

The range of assumptions tested for the Rasch model (3.8.2.1) was tested in the same way for the final LLTM. In the case that they were required, corrections to the model were made in the same way as for the Rasch models, and the new model retested. One additional test was for the absence of colinearity of fixed effects, which was not required for the Rasch models because their fixed effects were items and investigation of the LD assumption also addressed the colinearity assumption. In the case of LLTM, however, both assumptions must be tested. To assess colinearity between fixed effects, correlations between them were produced using R (R Core Team, 2014). The resulting matrices were examined for high correlations, which are a sign of colinearity. A figure of 0.866 was set as a cut off, as this is equal to a Variable Inflation Index (VIF) of 4. This figure is presented by Fox (2002) as being critical, as the confidence intervals are double the size of those for variables which are not correlated. The results of the testing of assumptions for the LLTMs are presented in 4.5.2.

3.8.3.1.2 Investigation of impact of attributes, subcomponents and components represented in the final model

In order to investigate the impact of attributes on test performance, it is useful to relate them to the theory which led to their specification (see 2.7). For this reason, attributes in the final model were grouped in categories according to the subcomponent and component to which they belonged. The combined influence (3.8.2.2.2) of each could then be calculated. These statistics were, however, based on the coefficients of indicators estimated independently, rather than in the final LLTM. This was done to avoid inaccuracy due to colinearity (3.8.2.2).

3.8.3.1.3 Strategy for assessing variance explained

LLTM cannot be expected to explain more variance than the Rasch model because it involves the decomposition of the item difficulty term in the model (see

2.9.3.3.3). For this reason, the results of the LLTM are often compared to the Rasch model for appraisal (De Boeck & Wilson, 2004a; Kubinger, 2009; Wilson & Moore, 2012). In other words, the variance explained by the Rasch model is the upper limit of what LLTM can be expected to explain. The lower limit is represented by an empty model, which is identical to the final LLTM but has not attribute-based fixed effects, although any fixed effects to account for LD are retained. The amount of variance explained by the final LLTM can be expressed as a point on this scale, therefore. Before the amount of variance explained was ascertained, however, the LRT was applied to determine that there was a significant difference between the following pairs of models:

- empty model and final LLTM
- final LLTM and the final Rasch model

A common approach to determining the comparative increase in variance explained by models such as LLTM is given in Equation 4. It is the difference between the figure for deviance[13] in the reference model and the improved model standardised to the scale of the reference model (the denominator). In order to express the difference in deviance for any two models as a value on the scale between the empty and corrected Rasch model, the denominator in Equation 4 becomes the deviance for the Rasch model subtracted from the empty model.

[13] Deviance is -2 x the log of the likelihood of the parameter estimates found during model estimation (Gelman & Hill, 2006). The natural log is used, as this facilitates simpler calculations (DeMars, 2010).

$$R_{\Delta}^{2} = \frac{(G_{R}^{2} - G_{F}^{2})}{G_{R}^{2}}$$

Where

G_{R}^{2} is the deviance of the reduced model

G_{F}^{2} is the deviance of the full model

Equation 4 Difference in R_{Δ}^{2} (De Ayala, 2009:141)

Results of the analysis of the amount of variance explained are found in 0.

3.9 Chapter summary

In this chapter, the method used to prepare and analyse the data has been set forth. It included the derivation of the required indicators, their individual analysis and selection for the final model, and the analysis of the final model itself. In addition, a range of descriptive statistics were specified: for the response data, the test materials and the information about candidate backgrounds. These are clearly ancillary but, in addition to presenting a picture of the data to be analysed, also provide information about the generalisability of the results of the main analysis. The next chapter will present the findings of the main analysis and the key findings of the descriptive statistics.

4 Results

4.1 Introduction

In this chapter, the results of the analysis described in 2.11 are presented. The chapter begins with the results of the testing of the sampling described in 3.3.2; descriptive statistics follow (3.4). The analysis of expert judgement to select relevant text for each item and the judgement required to construct four indicators are described next. The main analysis of the December 2005 FCE data began with the fitting of a Rasch model and the testing of its assumptions to determine the modifications needed for the Logistic Linear Test Model (LLTM) which followed. Analysis of the assumptions and the outcomes of its application to the data are described.

4.2 Validation and descriptive statistics of sample

Initial analysis involved calculating descriptive statistics for the response data and the test materials, as described in 3.4. These statistics provide an overview of the data and furnish evidence for the generalisability of the findings of the study. As discussed in 3.4, this is important because it shows that the results of the main analysis are also likely to be applicable to other forms of the same test. In order to conduct the main analysis, it was first necessary to sample the response data. Before the descriptive statistics were produced, the sample was validated. As described in 3.3.2, this was done to verify that the sampling did not introduce any distortion into the data which might affect later analyses. Results for both the validation of the sample and the descriptive statistics are presented in this section and in Appendix 4: Independent-Samples Wald-Wolfowitz Runs Test results, Appendix 5: summary of response matrices, Appendix 6: summary of the score matrices.

4.2.1 Validation of sample

The sampling of the data is described in 3.3.2.2, where the number of cases in the data was reduced from 25,803 to 9,961. Items were compared using the Independent-Samples Wald-Wolfowitz Runs Test (Wald and Wolfowitz, 1940).

This test compares the distributions of samples based on sequences of numbers. Unlike the T-Test, for example, nothing is assumed about the distribution of the variables, so the test is suitable for binary data, such as dichotomous item scores. An asymptotic significance of 1 was found, indicating that the null hypothesis was held: the distribution of scores across categories was not significantly different. The detailed results of the Wald-Wolfowitz Runs Test can also be found in Appendix 4: Independent-Samples Wald-Wolfowitz Runs Test results. Although the results of this test indicate that the sample is sufficiently representative of the data set from which it was drawn, frequency statistics for test responses and resulting scores were also calculated for both data sets. Scrutiny of these tables, which can be found in Appendix 8: descriptive statistics for candidate background data and Appendix 9: descriptive statistics for test materials, reveals, as with the results of the Wald-Wolfowitz Runs Test, that the responses in the sample closely reflect those in the parent data set.

4.2.2 Descriptive statistics

After the sample was validated, descriptive statistics were calculated. These were based on the sample data, the candidate background information and the test materials. As described in 3.4, their purpose was to provide an overview of the data.

4.2.2.1 Descriptive statistics of the measurement properties of the data

Statistics concerning the distribution of scores are contained in rows 1 – 8 of Table 31, with graphical representations of the distributions in Appendix 7: score distributions. Both statistics and graphical representations are given for each test part (task) and for the combination of all parts. The statistics presented in Table 31 are arranged with columns representing test parts and rows representing each statistic. The statistics are easiest to interpret as indicators of deviance from the normal distribution. In addition to helping to build up a picture of the data and the differences between the test parts, these statistics are helpful when interpreting parametric statistics, such as the standard deviation (row 5, Table 31). Parametric statistics assume a particular distribution, usually the normal distribution. Deviation from the assumed distribution is a sign that parametric statistics do not hold. For reference, the normal distribution is perfectly symmetrical, with a single

most common score (mode), which is identical to the mean and the median. Skew and kurtosis (an indicator of flatness or peakedness of the shape of the distribution) are both 0.

In all cases, it can be seen from Table 31 that the distributions of the test parts and all parts together are negatively-skewed, which means that scores are more frequent in the higher portion of the score range. In other words, the candidates found the items relatively easy overall. Although the scores are not normally distributed since they exhibit skew and kurtosis, they may still be considered to approximate to the normal distribution. Bachman (2004) suggests that, as a rule of thumb, if the skew and kurtosis indices are between -2 and 2, the distributions may be considered sufficiently normal to support parametric analyses. This is clearly the case with the data for the current study. It should also be noted, however, that despite all parts being approximately normal, the statistics indicated variation between the parts. For example, the Mean P shows that Part 1 was easiest overall (0.74), and Part 3 the most difficult (0.59).

Table 31 Descriptive statistics generated from the sample data for each test task

	Statistic or Index	Part 1	Part 2	Part 3	Part 4	All parts
1	**Mean Score**	5.18	5.64	4.11	8.98	23.91
2	**Median Score**	5	6	4	9	24
3	**Modal Score**	5	6	7	10	24
4	**Variance**	2.57	3.00	4.47	6.05	35.50
5	**Standard Deviation**	1.60	1.70	2.12	2.46	5.96
6	**Skew**	-0.81	-0.56	-0.06	-0.49	-0.33
7	**Kurtosis**	0.15	-0.24	-1.10	-0.20	-0.48
8	**Cronbach's Alpha**	0.66	0.52	0.74	0.63	0.83
9	**Standard Error of Measurement (SEM)**	0.93	1.20	1.08	1.50	2.47
10	**Mean P**	0.74	0.71	0.59	0.69	0.68
11	**Mean Item-Total**	0.58	0.48	0.62	0.43	0.38
12	**Mean Biserial Correlation**	0.84	0.64	0.80	0.59	0.52

4.2.2.2 Descriptive statistics for the candidate background characteristics - sample

Results for the analysis of candidate background characteristics can be found in Appendix 8: descriptive statistics for candidate background data. As with the other descriptive statistics, they provide an insight into the nature of the data used in the current study. One of the aims of the sampling was to reduce the influence of strongly-represented L1s because the linguistic influence of L1, or indeed more obscure factors related to educational or social similarities in countries with the same L1 might be expected to affect candidate response patterns. Scrutiny of the tables provided in Appendix 8: descriptive statistics for candidate background data, show that a more equal distributions of L1s was indeed achieved. As discussed in 3.2.2, it was expected that the restructuring of L1s would also affect some other characteristics, and this can also be seen. To determine the full effect of the sampling on the composition of background, descriptive statistics for the original crossed, cleaned data set is also provided.

A final addition to the statistics found in this section are a reference set of summary statistics, which represent the typical FCE candidature. Similarity between the reference set and the sample would show that the sample is representative of FCE more generally than just the December 2005 test form. This is important when considering whether conclusions drawn from the results of the main analysis may also apply to other forms of FCE. The reference statistics were published by Khalifa and Weir (2009), and calculated from reference data compiled from multiple sessions which took place in 2007. The representativeness of test materials is dealt with in 3.4.1 and 4.2.3. Only categories for which figures were reported by Khalifa and Weir (2009) were investigated and were consequently reported in this study.

4.2.3 Descriptive statistics for the test materials

In this section, descriptive statistics for the FCE Dec 05 test materials are contrasted with reference data provided by Khalifa and Weir (2009:76, 122, 131) and described in 0. As with candidate background data (4.2.2.2), the comparison is interesting because similarity provides evidence that the results of the main analysis are applicable to other forms of the same test. The table in Appendix 9:

descriptive statistics for test materials shows text length to be 34% longer overall in the texts for this study than the reference figure. According to Khalifa and Weir (2009), this is closer to the 3,000 word target for the Certificate in Advanced English (CAE), which is the C1 sister exam of FCE. The most likely reason for this is that, FCE underwent a revision before the publication of Khalifa and Weir (2009) in which Part 1 was removed. If this was done for the 2005 materials, the total number of words would be 1,911, much closer to that of the reference data.

As a result of this analysis, there are some indications that the reading texts for the 2005 data are broadly comparable, but perhaps marginally easier than those of the reference data. This suggests, together with statistics on candidate background characteristics, that data from the December 2005 FCE form was representative of FCE more generally.

4.3 Preparation for the main analysis

In order to prepare for the main analysis, it was necessary to extract information from the test materials to form indicator variables which make up the incidence matrix (3.7.1). As the incidence matrix is an item-by-indicator matrix, a value for each item on each indicator was necessary (see 3.7). A first step in this process was to demarcate the text which was particularly relevant to each item. Indices for each item was then derived from the text which was related to it. In most cases, this was done either by analysing the text automatically using software, or by applying rules manually. Only two indicators (holistic negation and fronted structures) were based on direct judgements concerning properties of the text. This section concerns the results of judgements of the relevant text for each item and the direct judgements for holistic negation and fronted sentences.

4.3.1 Expert judgement

4.3.1.1 Agreement in expert judgement of relevant text for each option

Three judges were asked to select the segments of text for each option which they felt contained the key information relating to each. Clearly, it is important that there should be a high level of agreement between the three judges, as this would increase confidence in the accuracy of their judgements. Table 32 is arranged with items in rows and, in columns two to ten, item options. The number of

judgemental agreements for each option of each item are contained in the cells. They are totalled in column eleven, the maximum possible number of agreements for that row is in the next column and, finally, the proportion of agreement is provided. Figures for agreement range from 0 (no agreement) to 1 (complete agreement between judges). Three items (items 11, 23 and 34) obtained complete agreement between the judges. The lowest figure proportion of agreement was 0.25 (item 29). Although it seems low, as Table 32 shows, only on option A was there no agreement; agreement between two of the three judges was present for the remaining three options. Furthermore, although a low agreement index implies a degree of uncertainty about at least some of the text selected, it does not indicate that text finally selected for the analysis is irrelevant. After judgements were collated and agreement indices calculated, a process to select text in cases of uncertainty was employed (3.5.1).

Table 32 Number of agreements per item option between three experts

	Option									Agreements	Total opportunities for agreement	Proportion of agreement
	A	B	C	D	E	F	G	H	I			
1	3	3	3	1	3	1	3	1	3	21	27	0.78
2	1	3	1	1	1	1	3	3	1	15	27	0.56
3	1	1	1	3	0	1	1	3	0	11	27	0.41
4	1	1	1	2	1	0	1	1	1	9	27	0.33
5	3	1	3	1	3	0	1	3	0	15	27	0.56
6	1	1	1	1	3	0	3	0	1	11	27	0.41
7	3	1	2	3	3	1	3	3	1	20	27	0.74
8	3	3	1	3						10	12	0.83
9	2	1	1	1						5	12	0.42
10	3	3	1	1						8	12	0.67
11	3	3	3	3						12	12	1.00
12	3	3	3	1						10	12	0.83
13	2	3	1	3						9	12	0.75
14	3	0	0	2						5	12	0.42
15	1	3	3	3						10	12	0.83
16	3	1	1	3	3	3	3	1	3	21	27	0.78
17	2	3	3	0	3	3	3	0	1	18	27	0.67
18	3	3	0	1	3	1	1	3	1	16	27	0.59
19	1	0	1	1	1	1	1	3	1	10	27	0.37
20	3	3	1	3	3	1	1	3	3	21	27	0.78
21	3	3	1	3	3	0	1	1	3	18	27	0.67
22	3	1	3	3	1	1	3	3	3	21	27	0.78
23	3	3	3	3						12	12	1.00
24	3	1	1	3						8	12	0.67
25	3	1	3	0						7	12	0.58
26	1	1	1	1						4	12	0.33
27	1	3	1	3						8	12	0.67
28	3	3	0	1						7	12	0.58
29	0	1	1	1						3	12	0.25
30	1	3	1	1						6	12	0.50
31	1	0	1	3						5	12	0.42
32	1	1	3	3						8	12	0.67
33	1	3	1	1						6	12	0.50
34	3	3	3	3						12	12	1.00
35	1	3	0	1						5	12	0.42

The proportion of agreement for judgements over each test part and all judgements were also calculated. Previous studies required judgements involving a series of categorical or ordered categories, and agreement was calculated either using agreement indices calculated from contingency tables, or by calculating the internal consistency within the data (Gwet, 2012). In the current study, there were no fixed categories, and this meant that it was impossible to calculate indices used in the field where investigation of rater agreement were required. Instead, agreement was defined as a match from a very wide range of possibilities: all the sentences in the text plus no response. For this reason, only the proportion of exact agreement was calculated. The results are contained in Table 33. No specific threshold was set in advance, due to lack of precedence. The figure of 0.61 for the agreement in the whole test represents around two thirds agreement and one third disagreement. This figure was accepted as adequate, as a figure indicating more agreement than disagreement, was expected.

When the proportion of agreement is calculated for each part, it can be seen that there is some variance in the level of agreement per part (Table 33). The highest levels of agreement were observed for Part 2. This may be attributable to the test method: items had only four options and the sequence of relevant text for each item followed the order of items.

Table 33 Number of agreements per item groupings between three experts

Sections concerned	Proportion of agreements
Overall	0.61
Part 1	0.54
Part 2	0.72
Part 3	0.66
Part 4	0.58

4.3.1.2 Agreement for judgements concerning holistic negation and fronted structures

Two judges were asked to review the test materials and select instances of holistic negation and fronted structures. These were defined following Weir (2013) (holistic negation) and Freedle and Kostin (1993) (fronted structures).. Agreement indices for direct judgements of holistic negation and fronted structures were calculated after the first round of judgements and are provided in Table 34. Agreements were defined as selection of the same text by both experts (columns two and three) and disagreements were defined as the selection of specific text by only one of the experts (columns four and five). The percentage of agreement per test part was calculated based on the total number of judgements for each attribute (columns six and seven) and is also presented in Table 34 (final two columns). For holistic negation was, this ranged from acceptable (Parts 1 and 2) to low (Part 3) to non-existent (Part 4). Those for fronted structures were, except for Part 4, non-existent. These figures were considered unacceptable, so as described in 3.6.1.7, judges were asked to discuss their choices and come to a mutual agreement for each indicator, which they subsequently did.

Table 34 Level of agreement in initial judgements of holistic negation and fronted structures

Part	Number of agreements		Number of disagreements		Total judgements		Agreements per total judgements	
	Holistic negation	Fronted structures	Holistic negation	Fronted structures	Holistic negation	Fronted structures	Holistic negation	Fronted structures
1	22	0	15	6	37	6	59.46%	0.00%
2	11	0	10	2	21	2	52.38%	0.00%
3	0	0	5	4	5	4	0.00%	0.00%
4	1	1	4	1	5	2	20.00%	50.00%
Total	34	1	34	13	68	14	50.00%	7.14%

4.4 Main analysis

The main analysis consisted of two stages: the testing of individual indicators, and the testing of a model composed of indicators found to be suitable. The results for each stage are given below.

4.4.1 Fitting of a Rasch model to the data

As discussed in 2.9.3.3.3 and 3.8.2.1, in a preliminary step in fitting LLTMs to the data, a Rasch model, with items as fixed effects, was fitted first and the assumptions of the model tested. The assumptions of unidimensionality and local item independence (LD) would be examined and corrections applied if violations found. After an acceptable model was found, the assumption of normality of person estimates were also tested. The final assumption mentioned in 2.9.3.3.4, that of the absence of colinearity is, for these Rasch models, identical to that of LD and, as determined in 3.8.2.1, was not tested. This is because, unlike LLTMs, the items (subject to the assumption of LD) and the fixed effects (subject to the assumption of absence of colinearity) are identical. Consequently, investigation of LD was assumed to cover both assumptions. At this point, it was possible to substitute items as fixed effects with item attributes as fixed effects and create a LLTM; this is discussed further in 2.9.3.3.30.

4.4.1.1 Assumptions of unidimensionality and LD

The first two models to be fitted were a unidimensional Rasch model and an empty unidimensional model, both without corrections of any kind. The comparative fit of the two models was tested using the Likelihood Ratio Test (LRT) to establish that the addition of the items made a significant improvement in fit. The results are contained in Table 35. After the titles, the first row contains information concerning the empty model (Model statistics) and the final row contains information about the unidimensional Rasch model. The statistics are, in order, the degrees of freedom of the model (parameters estimated); two fit indices (AIC and BIC), where lower numbers indicate better fit; the log likelihood and the deviance (the log likelihood multiplied by -2). The fit indices, log likelihood and deviance are explained in more detail in 3.8.2.1. They are all derived from the estimation procedure, Restricted Maximum Likelihood (REML) and are therefore tend to show a similar pattern when two models are compared. The final four columns contain statistics for the LRT. These include the input figures for the LTR: the absolute difference between the two figures for deviance (Chisq) and the absolute difference between the model degrees of freedom (Chi Df). The figure for the significance of the test, which is a chi-square test, is next (Pr(>Chisq)),

followed by a flag for the level of significance (key below the table). In the case of Table 35, the test shows that the addition of the items to the model make a significant difference, as would be expected. The unidimensional Rasch model is therefore preferred.

Table 35 LRT of two unidimensional models: the empty model and the Rasch model

	Model statistics					LRT			
	Df	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Chi Df**	**Pr(>Chisq)**	
Empty.uni	1	416885	416896	-208442	416883				
Rasch.uni	36	375063	375450	-187496	374991	41892	35	<2.2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

In order to investigate the assumption of unidimensionality, the model residuals were then examined for traces of secondary dimensions which were not accounted for by the model. This was done by scrutinising the scree plot of residuals given in Figure 7. The scree plot contains Eigenvalues, which are components into which residual variance is divided, ordered by magnitude from left to right. As recommended by DeMars (2010), the plot should be examined for a large drop between Eigenvalues, which would indicate that those points before the drop indicate significant secondary dimensions. In Figure 7, such a drop can be seen encompassing the first four points. In other words, there were four secondary dimensions which stood out in this analysis. As explained in 2.9.3.3.4 and 4.4.1, it was expected that there would be secondary dimensions related to the four test tasks and that these could be corrected by adding a dimension for each to the model.

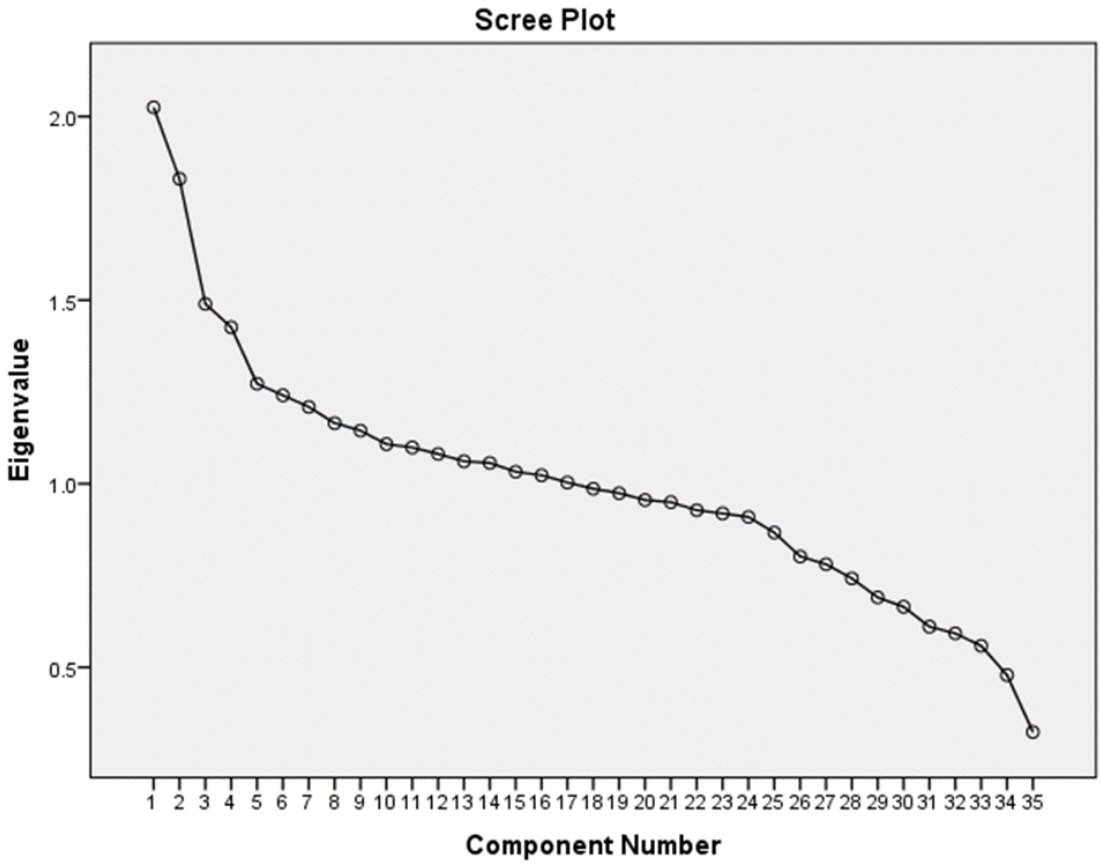


Figure 7 Scree plot, unidimensional Rasch model

A second model was specified, identical to the first model, but with a dimension corresponding to each test task. This model is derived from the RWLLTM (Rijmen & De Boeck, 2002), which relaxes the need for item characteristics to be fixed across all candidates (2.9.3.3.4). That the additional dimensions improved fit is shown by the LRT, Table 36. As the best fitting model was no longer unidimensional, the assumption for the model became that the dimensions specified would adequately account for the dimensional structure of the data. The scree plot for the residuals of this analysis is shown in Figure 8. The effect of adding the dimensions can be seen clearly by comparing it to the plot for the unidimensional model (Figure 7): the amount of variance explained by the most significant secondary dimensions (Eigenvalues) and the steepness of the drop between components on the scree plot were both reduced.

Table 36 LRT of two Rasch models: the unidimensional model and a model with four dimensions

Model statistics						LRT			
	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)	
Rasch.uni	36	375063	375450	-187496	374991				
Rasch.4d	45	369799	370283	-184855	369709	5282	9	<2.2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

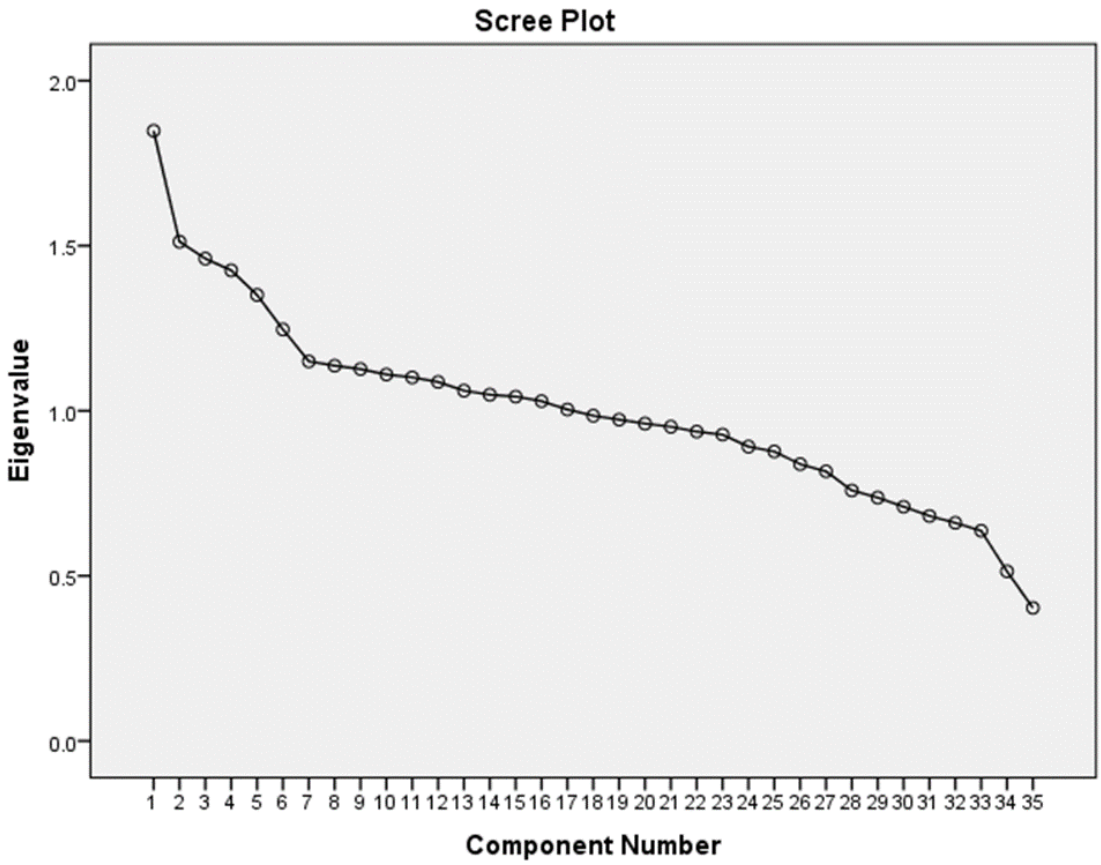


Figure 8 Scree plot, Rasch model with four dimensions

Since, the four expected dimensions were accounted for by the addition of a dimension for each, the LD assumption was addressed next. This was investigated using the squared Q_3 index (De Ayala, 2009), discussed in 3.8.2.1. The results are given in Table 37, Table 38,

Table 39 and Table 40. Each populated cell in the tables contains the square of the residual correlation of two items (listed across the top and down the left). For example, the figure for items 1 and 3 in Table 37 is 0.08. The square of the correlation coefficient *R* is R^2, which is the proportion of shared variance between two variables. The item pair 17 and 18 (Table 39) exhibit a particularly large residual correlation, resulting 34% of shared variance. For this reason, a dependency term to account for this and the model estimated again.

Table 37 Q_3^2 index, Rasch model with four dimensions, Part 1

	X1	***X2***	***X3***	***X4***	***X5***	***X6***	***X7***
X1	1.00						
X2	0.01	1.00					
X3	0.08	0.02	1.00				
X4	0.01	0.00	0.01	1.00			
X5	0.01	0.00	0.01	0.02	1.00		
X6	0.02	0.03	0.00	0.00	0.01	1.00	
X7	0.01	0.02	0.00	0.00	0.01	0.02	1.00

Table 38 Q_3^2 index, Rasch model with four dimensions, Part 2

	X8	***X9***	***X10***	***X11***	***X12***	***X13***	***X14***	***X15***
X8	1.00							
X9	0.00	1.00						
X10	0.00	0.00	1.00					
X11	0.00	0.00	0.00	1.00				
X12	0.00	0.00	0.00	0.00	1.00			
X13	0.00	0.00	0.00	0.00	0.00	1.00		
X14	0.00	0.00	0.00	0.00	0.00	0.00	1.00	
X15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 39 Q_3^2 index, Rasch model with four dimensions, Part 3

	X16	*X17*	*X18*	*X19*	*X20*	*X21*	*X22*
X16	1.00						
X17	0.02	1.00					
X18	0.02	0.34	1.00				
X19	0.01	0.02	0.03	1.00			
X20	0.07	0.01	0.02	0.01	1.00		
X21	0.01	0.04	0.03	0.00	0.01	1.00	
X22	0.00	0.01	0.01	0.00	0.01	0.03	1.00

Table 40 Q_3^2 index, Rasch model with four dimensions, Part 4

	X23	*X24*	*X25*	*X26*	*X27*	*X28*	*X29*	*X30*	*X31*	*X32*	*X33*	*X34*	*X35*
X23	1.00												
X24	0.00	1.00											
X25	0.02	0.00	1.00										
X26	0.00	0.02	0.00	1.00									
X27	0.00	0.00	0.00	0.00	1.00								
X28	0.00	0.00	0.00	0.02	0.00	1.00							
X29	0.00	0.00	0.01	0.00	0.01	0.00	1.00						
X30	0.01	0.00	0.00	0.00	0.00	0.00	0.00	1.00					
X31	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	1.00				
X32	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	1.00			
X33	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	1.00		
X34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	1.00	
X35	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

As dimensionality and LD are closely related (see 2.9.3.3.2), it was expected that corrections for dimensionality would have an effect on violations of LD and vice-versa. For this reason, when estimating a model with corrections for both assumptions, the scree plot (Figure 9) and the Q_3^2 index for the new model (Table 42, Table 43, Table 44 and Table 45) were scrutinised. First, however, a LRT was conducted (Table 41), and this showed that the model with the correction for the violation of LD fitted significantly better than its predecessor. It can be seen in Figure 9 that eigenvalues are yet lower. Furthermore, the drop observed between the first and second Eigenvalues in Figure 8 is no longer present, suggesting a reduction in the influence of LD.

Table 41 LRT of two Rasch models with four dimensions: one without any corrections for LD, and one with a correction for dependency between items 17 and 18.

Model statistics						LRT			
	Df	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Chi Df**	**Pr(>Chisq)**	
Rasch.4d	45	369799	370283	-184855	369709				
Rasch.4d.dep5	46	342802	343295	-171355	342710	26999	1	<2.2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

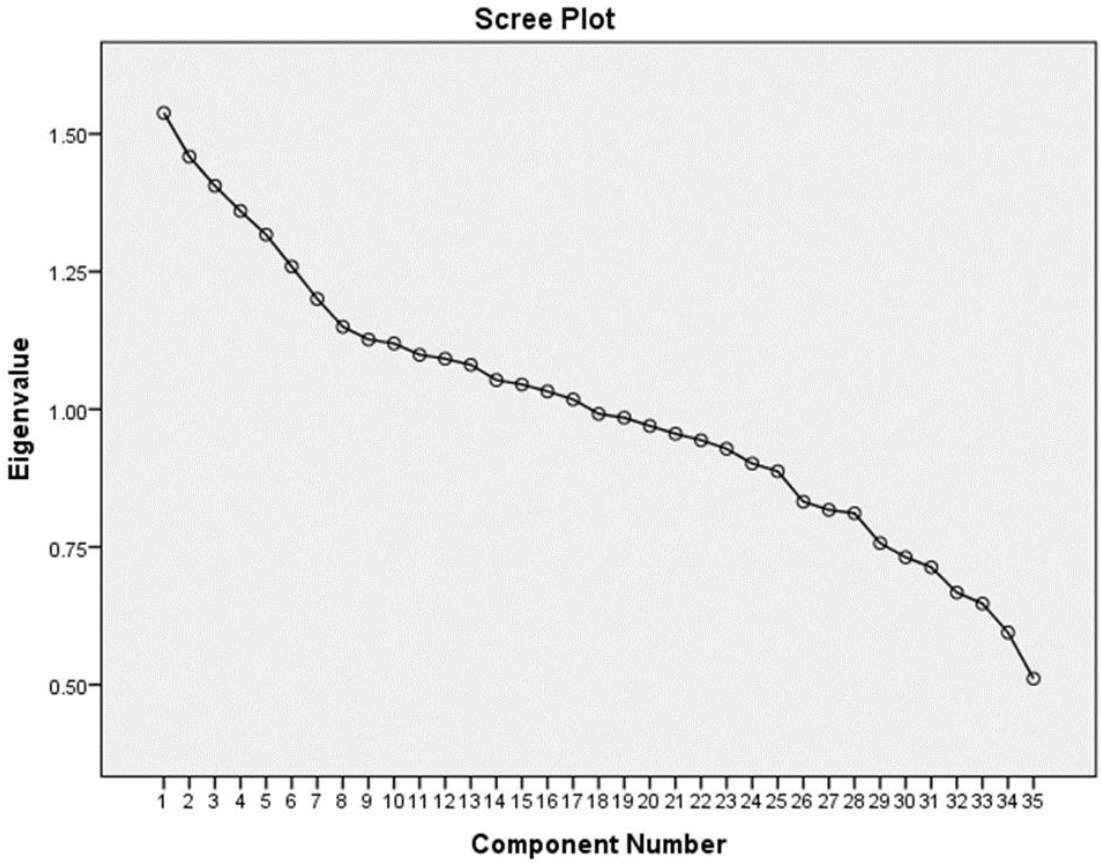


Figure 9 Scree plot, Rasch model with four dimensions and correction for dependency between items 17 and 18

The item pair with the largest Q_3^2 index remains that of items 17 and 18 (Table 44). But the value of 11% is much reduced from the previous 34%. For this reason, and because the scree plot in Figure 9 was acceptable, it was decided to accept the current model as a basis for the LLTMs, provided the final assumption could be satisfied. The assumption of a normal distribution for candidate ability estimates is dealt with in 4.4.1.2.

Table 42 Q_3^2 index, Rasch model with four dimensions and correction for dependency between items 17 and 18, Part 1

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
X1	1.00						
X2	0.01	1.00					
X3	0.09	0.02	1.00				
X4	0.01	0.00	0.01	1.00			
X5	0.01	0.00	0.00	0.02	1.00		
X6	0.01	0.02	0.00	0.00	0.02	1.00	
X7	0.01	0.02	0.00	0.00	0.01	0.03	1.00

Table 43 Q_3^2 index, Rasch model with four dimensions and correction for dependency between items 17 and 18, Part 2

	X8	*X9*	*X10*	*X11*	*X12*	*X13*	*X14*	*X15*
X8	1.00							
X9	0.00	1.00						
X10	0.00	0.00	1.00					
X11	0.00	0.00	0.00	1.00				
X12	0.00	0.00	0.00	0.00	1.00			
X13	0.00	0.00	0.00	0.00	0.00	1.00		
X14	0.00	0.00	0.00	0.00	0.00	0.00	1.00	
X15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 44 Q_3^2 index, Rasch model with four dimensions and correction for dependency between items 17 and 18, Part 3

	X16	*X17*	*X18*	*X19*	*X20*	*X21*	*X22*
X16	1.00						
X17	0.00	1.00					
X18	0.00	0.11	1.00				
X19	0.00	0.00	0.01	1.00			
X20	0.08	0.00	0.00	0.01	1.00		
X21	0.00	0.01	0.00	0.00	0.01	1.00	
X22	0.00	0.00	0.00	0.01	0.00	0.04	1.00

Table 45 Q_3^2 index, Rasch model with four dimensions and correction for dependency between items 17 and 18, Part 4

	X23	*X24*	*X25*	*X26*	*X27*	*X28*	*X29*	*X30*	*X31*	*X32*	*X33*	*X34*	*X35*
X23	1.00												
X24	0.00	1.00											
X25	0.02	0.00	1.00										
X26	0.00	0.02	0.00	1.00									
X27	0.00	0.00	0.00	0.00	1.00								
X28	0.00	0.00	0.00	0.02	0.00	1.00							
X29	0.00	0.00	0.01	0.00	0.01	0.00	1.00						
X30	0.01	0.00	0.00	0.00	0.00	0.00	0.00	1.00					
X31	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	1.00				
X32	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	1.00			
X33	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	1.00		
X34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	1.00	
X35	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

4.4.1.2 Assumption of normal distribution for candidate ability estimates

The normality of ability estimates was investigated by calculating descriptive statistics for the distribution for each dimension. These statistics are given in Table 46. In all cases, it can be seen that the mean and median are both very close to each other and to zero. Skew and kurtosis are within the +/-2 range (Bachman, 2004). For these reasons, it was concluded that the distribution of the estimates was, in the case of each dimension, sufficiently normal and did not constitute a serious violation of this assumption.

Table 46 Summary statistics for candidate ability estimates for each dimension of the Rasch model with four dimensions and correction of dependency between items 17 and 18

	Part 1	Part 2	Part 3	Part 4
Mean	-0.01	0.00	0.01	0.00
Median	0.01	-0.01	-0.07	0.00
SD	1.15	0.71	1.02	0.69
Variance	1.32	0.50	1.05	0.47
Max	2.24	1.63	2.16	1.59
Min	-3.77	-2.21	-3.00	-2.45
Range	6.01	3.83	5.16	4.03
Skew	-0.25	0.00	0.06	-0.02
Kurtosis	-0.43	-0.42	-0.72	-0.40

4.4.2 Results of analysis of indicators

4.4.2.1 Analysis of each indicator

As described in 3.8.2.2, indicators were analysed individually to provide accurate information about each and to determine which should be included in the final model. This involved determining whether the estimate for each complied with theoretical explanations, and whether the estimate was significant. The importance of the former is worth emphasising, as, in many fields, parsimony for theoretical, not only statistical, reasons is considered essential when modelling data statistically (Chou & Huh, 2012). The tables in this section contain information on each indicator tested, comprising:

- its effect on the chances of success (columns two and three in the tables, expressed in logs odds and as probability)
- the size of the statistical error associated with the estimates (column four)
- a z value (column 5), its p statistic (column 6) and a flag for significance (final column)

For indicators measured by continuous variables, positive log odds and probability values higher than 0.5 indicate an increased chance of success on an item containing the attribute represented by the indicator, or, in other words, an attribute which makes items easier. Lower values represent a decreased chance of success, or an attribute which makes items more difficult. Coefficients are provided in three different formats: in the form of log odds of success on the item, as the probability, and as a percentage, representing the influence (positive or negative) of the indicator on success on the item. The expected effect for the log odds is recorded in the second column, so that the coefficients can be more easily interpreted.

Factor-based, or categorical, indicators, are slightly different from continuous indicators, and estimates must be judged according to the way in which the levels of the indicator compare to each other. If an attribute is expected to increase item difficulty, the estimated coefficients for higher levels of the indicator would be expected to be smaller than those at lower levels. For attributes which decrease

item difficulty, higher levels should be larger than lower levels. Taken together, the levels of the indicator should form a monotonically increasing or decreasing set of estimates. The expected direction of change is indicated by the words 'harder' and 'easier' at either end of the scale, with a larger coefficient indicating an easier level. Estimates for levels which do not follow the sequence are a sign of a problematic indicator. In addition, the coefficient for each level must be statistically separable from those of other levels. In other words, the confidence intervals (twice measurement error) must not overlap – this is indicated in the text.

4.4.2.1.1 OP indicators

The OP indicators relating to the word recognition and lexical access aspects of the composite model (2.7.1) are provided in Table 47. Each row in the table contains values for one indicator (continuous indicators) or for one level of a factor-based indicator. For continuous indicators, the second column shows the expected direction of the coefficient, which follows in the form of log-odds in the next column. For factor-based indicators, the direction of the expected progression is indicated by the words 'harder' at the end where the smallest coefficient should be found, and 'easier' at the place for the largest coefficient. The figure for log-odds was converted into two other forms to facilitate its interpretation. First probability, which is a number between 0 and 1, with 0.5 indicating neither greater nor lesser chances of success on items. Influence is probability centred on 0 instead of 0.5, so that any negative number indicates an attribute which increases item difficulty, and any positive number indicates an attribute which makes the item easier. For the scale of any factor-based indicators, a zero point is set where absence for the level where absence of influence is expected. Figures for other levels may then be seen as relative to that zero point. For example, in Table 47, the zero point for the second indicator (X002.OP.BNC) is set for level 1, as words in this category are expected to increase difficulty the least. The standard error of the coefficient and a test of significance make up the last four columns of each table. The significance test compares the size of the coefficient to its error. If the

former is larger than two times the error, the value is significant to the probability in the penultimate column and flagged in the final column.

Among the continuous-based indicators, X006.OP.CELEX.cont.f (content word frequency) and its log (X008.OP.CELEX.cont.log) match the expected direction of influence – the estimates are positive and, therefore, more frequent words (higher indices) make the items easier. X010.OP.hypernymy was also retained for the same reasons.

Of the indicators for which the direction of influence did not match expectations, the estimate for the number of syllables per word (X001.OP.syll) had the largest magnitude. This was not taken to represent counter-evidence to current theories on the impact of the number of syllables on cognitive processing (Weir, 2013), however, as this study focussed on the construct of the test in terms of the theory and not the opposite. Since the results of the analysis depend on the possibility to interpret them in terms of prevailing theory, any which were found to be uninterpretable were dropped from the study. Their status was not understood to be an indication about the nature of the test, or of theory. This is important, as several explanations for the unexpected figures could have been argued. In addition to aberrant theory or a poor test, capitalisation on chance, or large measurement error generated in the process could have rendered the indicator unrepresentative of what it purports to measure. It was beyond the scope of the current study to indubitably determine a cause, however. A parsimonious approach was, therefore, followed, whereby indicators with uninterpretable coefficients are left out (see 3.8.2.2).

Unlike the CELEX indicators, higher numbers for the BNC and AWL indicators represent less frequent words. For this reason, higher log odds estimates are expected at lower levels of each. In neither case, however, did the values presented in Table 47 feature a monotonic decrease from the lowest level to the highest. For X002.OP.BNC, it was decided to collapse levels 1 to 3 to see if the new indicator would yield interpretable estimates. In essence, this change meant the

testing of the hypothesis that the BNC frequency of 1, 2, or 3k level words affected difficulty in approximately the same way as each other, whereas those of other levels had progressively increased difficulty. The levels of the AWL indicator were not collapsed in the same way, as the third level (15) corresponded exactly with items in Part 1 of the test. Collapsing the first two levels would leave an indicator which simply represented all differences between Part 1 and the other three parts, rather than providing evidence relating specifically to academic words. For this reason, AWL was dropped at this point.

Table 47 Estimates for OP word recognition and lexical access indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X001.OP.syll	-ve	0.58	0.64	0.14	0.02	26.83	<2e-16	***
X002.OP.BNC1	easier	1.70	0.85	0.00	0.03	58.94	<2e-16	***
X002.OP.BNC2		1.07	0.74	-0.10	0.01	88.98	<2e-16	***
X002.OP.BNC3		0.71	0.67	-0.17	0.01	64.36	<2e-16	***
X002.OP.BNC8		1.41	0.80	-0.04	0.02	77.62	<2e-16	***
X002.OP.BNC10		1.29	0.78	-0.06	0.03	47.52	<2e-16	***
X002.OP.BNC14		1.29	0.78	-0.06	0.02	85.36	<2e-16	***
X002.OP.BNC16		0.55	0.63	-0.21	0.02	22.34	<2e-16	***
X002.OP.BNC26	harder	0.39	0.60	-0.25	0.02	22.69	<2e-16	***
X003.OP.AWL3	easier	1.04	0.74	0.00	0.01	102.22	<2e-16	***
X003.OP.AWL4		0.84	0.70	-0.04	0.01	68.53	<2e-16	***
X003.OP.AWL15	harder	1.33	0.79	0.05	0.02	85.02	<2e-16	***
X006.OP.CELEX.cont.f	+ve	2.30	0.91	0.41	0.04	51.96	<2e-16	***
X007.OP.CELEX.all.f.log	+ve	-0.65	0.34	-0.16	0.03	-22.95	<2e-16	***
X008.OP.CELEX.cont.log	+ve	0.28	0.57	0.07	0.01	37.33	<2e-16	***
X010.OP.hypernymy	-ve	-0.05	0.49	-0.01	0.01	-5.89	0.00	***
X011.OP.polysemy	-ve	0.24	0.56	0.06	0.00	52.09	<2e-16	***
X012.OP.lex.density	-ve	0.00	0.50	0.00	0.00	9.27	<2e-16	***
X013.OP.concrete	+ve	0.00	0.50	0.00	0.00	-18.77	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

The results of collapsing the levels of the OP BNC indicator are given in Table 48. There it can be seen that, even with the modification to the indicator in this table, the coefficients for the levels in the third column do not display a monotonic progression. As a result of the analysis for the OP indicators for word recognition

and lexical access, only CELEX content word frequency (X006.OP.CELEX.cont.f) and its log counterpart (X007.OP.CELEX.cont.log) were retained.

Table 48 Estimates for OP BNC indicator with collapsed levels

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X002.OP.BNC.CLPS1LOW	easier	0.91	0.71	0.00	0.01	94.66	<2e-16	***
X002.OP.BNC.CLPS8		1.40	0.80	0.09	0.02	77.38	<2e-16	***
X002.OP.BNC.CLPS10		1.29	0.78	0.07	0.03	47.73	<2e-16	***
X002.OP.BNC.CLPS14		1.28	0.78	0.07	0.01	85.24	<2e-16	***
X002.OP.BNC.CLPS16		0.53	0.63	-0.08	0.02	21.91	<2e-16	***
X002.OP.BNC.CLPS26	harder	0.37	0.59	-0.12	0.02	21.58	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

The two indicators associated with syntactic parsing both had negative log odds estimates as expected and both were statistically significant (Table 49). The impact of both was, however, slight, with around 9% and 4% decreased chances of success respectively. Both indicators were retained.

Table 49 Estimates for OP syntactic parsing indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X014.OP.mod.noun	-ve	-0.35	0.41	-0.09	0.01	-26.66	<2e-16	***
X015.OP.left.emb	-ve	-0.17	0.46	-0.04	0.01	-22.45	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

The estimates for indicators associated with establishing propositional meaning are presented in

Table 50. Of the indicators derived from factors, only that representing fronted structures (X018.OP.fronted), is as predicted by theory. In other words, those items without fronted structures (level 1NO) are easier than those with fronted structures (level 2YES). The other two indicators, for holistic negation (X017.OP.hol.neg) and the number of propositions (X022.OP.props), did not exhibit a clear trend across their estimated coefficients. In the case of the indicator for propositions (X022.OP.props), examination of its thirteen levels did not reveal a clear way in which the categories might be usefully collapsed. F. In addition, the

indicator for proposition density (X000.OP.prop.dens) had a positive coefficient when a negative one was expected. As a result, both indicators were dropped along with the indicator for holistic negation (X017.OP.hol.neg). The passive voice (X019.OP.passive) has a very slight impact (0.04%) but was in the direction implied by theory, so is retained.

Table 50 Estimates for OP establishing propositional meaning indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X016.OP.neg	-ve	0.00	0.50	0.00	0.00	2.21	0.03	*
X017.OP.hol.neg0	easier	0.92	0.72	0.00	0.01	96.20	<2e-16	***
X017.OP.hol.neg1		0.58	0.64	-0.08	0.02	35.00	<2e-16	***
X017.OP.hol.neg2		1.18	0.77	0.05	0.02	78.04	<2e-16	***
X017.OP.hol.neg6	harder	1.29	0.78	0.07	0.02	85.65	<2e-16	***
X018.OP.fronted1NO	easier	1.04	0.74	0.00	0.01	109.01	<2e-16	***
X018.OP.fronted2YES	harder	0.39	0.60	-0.14	0.02	22.15	<2e-16	***
X019.OP.passive	-ve	0.00	0.50	0.00	0.00	-6.63	0.00	***
X022.OP.props5	easier	1.13	0.76	0.00	0.02	70.61	<2e-16	***
X022.OP.props6		0.83	0.70	-0.06	0.02	53.87	<2e-16	***
X022.OP.props7		1.45	0.81	0.05	0.02	85.49	<2e-16	***
X022.OP.props8		0.40	0.60	-0.16	0.02	23.26	<2e-16	***
X022.OP.props9		0.74	0.68	-0.08	0.02	41.76	<2e-16	***
X022.OP.props20		0.68	0.66	-0.09	0.02	28.20	<2e-16	***
X022.OP.props30		1.06	0.74	-0.01	0.02	64.46	<2e-16	***
X022.OP.props39		1.65	0.84	0.08	0.03	57.16	<2e-16	***
X022.OP.props41		1.26	0.78	0.02	0.03	47.72	<2e-16	***
X022.OP.props43		1.02	0.73	-0.02	0.03	40.17	<2e-16	***
X022.OP.props48		0.52	0.63	-0.13	0.02	21.74	<2e-16	***
X022.OP.props63		1.27	0.78	0.02	0.01	89.84	<2e-16	***
X022.OP.props84	harder	0.37	0.59	-0.16	0.02	22.29	<2e-16	***
X000.OP.prop.dens	-ve	0.66	0.66	0.16	0.06	10.22	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

4.4.2.1.2 SEARCH indicators

The estimates for indicators associated with the search component are contained in Table 51. Among these coefficients, that for the match between the OP text and the relevant text for READ (X053.SEARCH.LSA.term) was not significant and was therefore rejected. The indicator for demarcation of the relevant reading text for an item (X052.SEARCH.demarc) did not display a monotonic pattern. However, it

was decided to re-specify the indicator by collapsing categories 0 and 1 together and 2 and 3 together. Category 0 represented no demarcation whatsoever; category 1 an approximately matched area of text. For example, in Part 3, the location of the gap to be filled is clear, but not the precise location of the crucial text in the reading passage which will determine the choice of option. Category 2 represents a precisely-demarcated paragraph (as with Part 1) and category 3 is yet more precise demarcation, such as a specific line. Collapsing categories in the way suggested, therefore, represents a test of the hypothesis that the kind of demarcation in Part 3 does not help, demarcation with precise boundaries makes the item easier. Finally, the indicator representing items following the same order as the relevant information in the passage (X051.SEARCH.order) was significant and met theoretical expectations, so was retained.

Table 51 Estimates for SEARCH indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X051.SEARCH.order1NO	harder	1.02	0.74	0.00	0.01	94.84	<2e-16	***
X051.SEARCH.order2YES	easier	1.10	0.75	0.02	0.01	96.68	<2e-16	***
X052.SEARCH.demarc0	harder	1.03	0.74	0.00	0.01	101.75	<2e-16	***
X052.SEARCH.demarc1		0.38	0.59	-0.07	0.01	21.94	<2e-16	***
X052.SEARCH.demarc2		1.11	0.75	0.09	0.01	86.11	<2e-16	***
X052.SEARCH.demarc3	easier	0.69	0.67	0.00	0.01	28.25	<2e-16	***
X053.SEARCH.LSA.term	+ve	-0.11	0.47	-0.03	0.07	-1.43	0.15	

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

The results for the indicator for demarcatedness with collapsed categories are given in Table 52. This indicator was significant and accorded with theory, although the effect was slight. This indicator was retained for this component, along with the others mentioned above.

Table 52 Estimates for SEARCH demarcatedness indicator with collapsed levels

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X052.SEARCH.demarc.CLPS1LOW	harder	1.04	0.74	0.00	0.01	100.30	<2e-16	***
X052.SEARCH.demarc.CLPS2HIGH	easier	1.10	0.75	0.01	0.01	89.00	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

4.4.2.1.3 READ indicators

The estimated coefficients for the indicators associated with word recognition and lexical access for the read component are displayed in Table 53. Estimators for all indicators were statistically significant. Among those measured by continuous variables, only those for the number of syllables (X026.READ.syll), CELEX frequency of content words (X031.READ.CELEX.cont.f), the log of the CELEX frequency for all words (X032.READ.CELEX.all.f.log), hypernymy (X035.READ.hypernymy) and lexical density (X037.READ.lex.density) had the predicted direction of impact on items, and were therefore retained.

Table 53 Estimates for READ word recognition and lexical access indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X026.READ.syll	-ve	-0.97	0.28	-0.22	0.03	-37.51	<2e-16	***
X027.READ.BNC3	easier	1.74	0.85	0.00	0.02	95.31	<2e-16	***
X027.READ.BNC4		0.42	0.60	-0.25	0.02	26.01	<2e-16	***
X027.READ.BNC6		1.81	0.86	0.01	0.02	108.54	<2e-16	***
X027.READ.BNC8		0.90	0.71	-0.14	0.02	36.30	<2e-16	***
X027.READ.BNC10		1.25	0.78	-0.07	0.03	47.54	<2e-16	***
X027.READ.BNC11		0.58	0.64	-0.21	0.02	27.11	<2e-16	***
X027.READ.BNC16		0.76	0.68	-0.17	0.01	72.58	<2e-16	***
X027.READ.BNC17		1.67	0.84	-0.01	0.03	57.81	<2e-16	***
X027.READ.BNC21		0.68	0.66	-0.19	0.02	30.04	<2e-16	***
X027.READ.BNC26	harder	0.80	0.69	-0.16	0.01	64.43	<2e-16	***
X028.READ.AWL0	easier	1.68	0.84	0.00	0.03	56.02	<2e-16	***
X028.READ.AWL4	harder	1.07	0.74	-0.10	0.01	109.24	<2e-16	***
X031.READ.CELEX.cont.f	+ve	0.77	0.68	0.18	0.03	27.19	<2e-16	***
X032.READ.CELEX.all.f.log	+ve	0.72	0.67	0.17	0.04	20.39	<2e-16	***
X033.READ.CELEX.cont.log	+ve	-0.34	0.42	-0.08	0.01	-43.20	<2e-16	***
X035.READ.hypernymy	-ve	-0.09	0.48	-0.02	0.02	-6.01	0.00	***
X036.READ.polysemy	-ve	0.43	0.61	0.11	0.01	65.51	<2e-16	***
X037.READ.lex.density	-ve	-0.02	0.50	0.00	0.00	-17.46	<2e-16	***
X038.READ.concrete	+ve	0.00	0.50	0.00	0.00	-3.44	0.00	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

When the indicators based on factors were considered, only the AWL indicator (X028.READ.AWL) corresponded to theoretical expectations, in that the level with

less frequent words related to increased item difficulty. The BNC indicator (X027.READ.BNC) did not present a clear pattern, but categories were collapsed into words in low (levels 3 to 11) and high (levels 16 to 26) levels, and the new indicator analysed. The results are provided in Table 54. They show that words in higher levels indeed made items more difficult. Both the AWL indicator, and the BNC indicator with collapsed categories were retained.

Table 54 Estimates for READ BNC indicator with collapsed levels

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X027.READ.BNC.CLPS1LOW	easier	1.31	0.79	0.00	0.01	115.82	<2e-16	***
X027.READ.BNC.CLPS3HIGH	harder	0.92	0.72	-0.07	0.01	91.62	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

Although both significant, neither of the indicators associated with syntactic parsing produced estimates which matched expectations (Table 55). For this reason, neither was retained.

Table 55 Estimates for READ syntactic parsing indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X039.READ.mod.noun	-ve	0.41	0.60	0.10	0.02	25.18	<2e-16	***
X040.READ.left.emb	-ve	0.07	0.52	0.02	0.00	45.35	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

The estimates for indicators related to establishing propositional meaning are contained in Table 56 and Table 57. Among these indicators, only that for the passive voice (X044.READ.passive) was retained, although its influence was weak and its significance low. The remaining continuous indicators, although significant, did not accord with theory and were, therefore, dropped.

Among the indicators derived from factors, none was theoretically interpretable as they stood, so it was decided to collapse categories within each to create new indicators to be analysed. For X042.READ.hol.neg, instead of counting the number of instances, it was decided to form two groups: items with one or more instances

('YES'), and items with none ('NO'). For fronted structures (X043.READ.fronted), reading text with one or two instances were reclassified as 'LOW', those with more instances as 'HIGH', and those without instances as 'NONE'. Despite not offering a clear pattern amongst the estimates for each level, the indicator counting propositions was also reconfigured because propositions in the reading passage were considered important in other studies (e.g. Embretson & Wetzel, 1987). It was divided into three groups: 'LOW', 'MID' and 'HIGH', the threshold being somewhat arbitrary (50 and 100).

Table 56 Estimates for READ establishing propositional meaning indicators I

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X041.READ.neg	-ve	0.01	0.50	0.00	0.00	13.96	<2e-16	***
X042.READ.hol.neg0	easier	0.90	0.71	0.00	0.01	88.18	<2e-16	***
X042.READ.hol.neg1		2.03	0.88	0.17	0.02	104.91	<2e-16	***
X042.READ.hol.neg2		1.24	0.78	0.06	0.01	84.25	<2e-16	***
X042.READ.hol.neg3		0.17	0.54	-0.17	0.02	9.10	<2e-16	***
X042.READ.hol.neg4		0.16	0.54	-0.17	0.02	8.81	<2e-16	***
X042.READ.hol.neg5		0.77	0.68	-0.03	0.02	39.68	<2e-16	***
X042.READ.hol.neg6		1.96	0.88	0.16	0.03	59.94	<2e-16	***
X042.READ.hol.neg8		0.79	0.69	-0.02	0.02	40.38	<2e-16	***
X042.READ.hol.neg12	harder	2.30	0.91	0.20	0.04	63.50	<2e-16	***
X043.READ.fronted0	easier	1.22	0.77	0.00	0.01	107.71	<2e-16	***
X043.READ.fronted1		0.92	0.71	-0.06	0.01	77.06	<2e-16	***
X043.READ.fronted2		1.09	0.75	-0.02	0.03	42.82	<2e-16	***
X043.READ.fronted3		1.77	0.85	0.08	0.03	57.72	<2e-16	***
X043.READ.fronted4	harder	0.24	0.56	-0.21	0.02	13.88	<2e-16	***
X044.READ.passive	-ve	0.00	0.50	0.00	0.00	-2.28	0.02	*

Table 57 Estimates for READ establishing propositional meaning indicators II

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X047.READ.props35	easier	0.44	0.61	0.00	0.02	18.57	< 2e-16	***
X047.READ.props37		1.27	0.78	0.17	0.03	47.92	< 2e-16	***
X047.READ.props41		2.38	0.92	0.31	0.04	65.62	< 2e-16	***
X047.READ.props50		1.18	0.77	0.16	0.02	70.70	< 2e-16	***
X047.READ.props53		1.92	0.87	0.26	0.03	61.44	< 2e-16	***
X047.READ.props58		0.46	0.61	0.01	0.02	19.66	< 2e-16	***
X047.READ.props60		0.97	0.73	0.12	0.02	54.07	< 2e-16	***
X047.READ.props61		1.00	0.73	0.12	0.03	39.53	< 2e-16	***
X047.READ.props63		-0.30	0.42	-0.18	0.02	-13.03	< 2e-16	***
X047.READ.props64		1.18	0.76	0.16	0.03	45.30	< 2e-16	***
X047.READ.props65		1.67	0.84	0.23	0.03	57.11	< 2e-16	***
X047.READ.props66		0.64	0.65	0.05	0.02	26.65	< 2e-16	***
X047.READ.props70		-0.04	0.49	-0.12	0.02	-1.54	0.12	
X047.READ.props76		0.37	0.59	-0.02	0.03	13.76	< 2e-16	***
X047.READ.props77		1.07	0.74	0.14	0.03	41.98	< 2e-16	***
X047.READ.props80		2.25	0.90	0.30	0.04	62.74	< 2e-16	***
X047.READ.props82		0.55	0.63	0.03	0.02	22.54	< 2e-16	***
X047.READ.props86		1.42	0.81	0.20	0.03	51.88	< 2e-16	***
X047.READ.props87		-0.53	0.37	-0.24	0.03	-19.03	< 2e-16	***
X047.READ.props95		0.69	0.67	0.06	0.02	28.00	< 2e-16	***
X047.READ.props97		1.63	0.84	0.23	0.03	56.43	< 2e-16	***
X047.READ.props98		1.66	0.84	0.23	0.03	52.10	< 2e-16	***
X047.READ.props99		-0.05	0.49	-0.12	0.03	-1.90	0.06	.
X047.READ.props100		2.93	0.95	0.34	0.04	68.50	< 2e-16	***
X047.READ.props106		0.85	0.70	0.09	0.03	33.97	< 2e-16	***
X047.READ.props131		1.25	0.78	0.17	0.03	43.76	< 2e-16	***
X047.READ.props136		-1.38	0.20	-0.41	0.04	-33.56	< 2e-16	***
X047.READ.props140		2.59	0.93	0.32	0.04	66.52	< 2e-16	***
X047.READ.props162		-1.96	0.12	-0.48	0.05	-41.37	< 2e-16	***
X047.READ.props196		0.13	0.53	-0.08	0.03	4.60	0.00	***
X047.READ.props197	harder	2.07	0.89	0.28	0.03	60.13	< 2e-16	***
X000.READ.prop.dens	-ve	0.66	0.66	0.16	0.06	10.22	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

The estimates for these newly constructed indicators are presented in Table 58. They show that only the indicator for fronted structures with collapsed levels concords with theory. For this reason, it was retained.

Table 58 Estimates for READ holistic negation, fronted and propositions indicators with collapsed levels

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X042.READ.hol.neg.CLPSNO	easier	0.97	0.73	0.00	0.01	92.33	<2e-16	***
X042.READ.hol.neg.CLPSYES	harder	1.12	0.75	0.03	0.01	107.97	<2e-16	***
X043.READ.fronted.CLPS0NONE	easier	1.25	0.78	0.00	0.01	109.21	<2e-16	***
X043.READ.fronted.CLPS1LOW		0.95	0.72	-0.06	0.01	80.35	<2e-16	***
X043.READ.fronted.CLPS3HIGH	harder	0.60	0.65	-0.13	0.02	37.72	<2e-16	***
X047.READ.props.CLPS1LOW	easier	1.34	0.79	0.00	0.02	78.57	<2e-16	***
X047.READ.props.CLPS2MID		0.99	0.73	-0.06	0.01	99.67	<2e-16	***
X047.READ.props.CLPS3HIGH	harder	1.15	0.76	-0.03	0.01	85.09	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

Estimates for indicators associated with establishing a coherent textbase are presented in Table 59. All estimates were statistically significant but only two were readily interpretable: the indicator representing connectives (X045.READ.connect) and stem overlap (X046.READ.stem.o). These were therefore retained. The indicator for sentences (X000.READ.sentence) showed some signs of a pattern, with the last few levels having negative coefficients. Levels were not collapsed for this indicator, however, as middle levels (9 to 16) have, on average, higher coefficients than those of the levels below them. In other words, if this indicator were collapsed into three categories, it seems likely that the middle levels would have appeared easier than the lower levels, which represented fewer sentences. For this reason, the initial indicator was dropped.

Table 59 Estimates for READ establishing a coherent textbase indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X045.READ.connect	+ve	0.00	0.50	0.00	0.00	10.03	<2e-16	***
X046.READ.stem.o	+ve	0.27	0.57	0.07	0.03	9.30	<2e-16	***
X034.READ.type.tok	-ve	1.20	0.77	0.27	0.18	6.69	0.00	***
X000.READ.sentence4	easier	1.26	0.78	0.00	0.01	84.99	<2e-16	***
X000.READ.sentence5		1.14	0.76	-0.02	0.01	80.31	<2e-16	***
X000.READ.sentence6		1.29	0.78	0.00	0.02	76.23	<2e-16	***
X000.READ.sentence7		0.76	0.68	-0.10	0.02	50.53	<2e-16	***
X000.READ.sentence8		0.87	0.70	-0.08	0.02	53.18	<2e-16	***
X000.READ.sentence9		1.85	0.86	0.08	0.02	84.43	<2e-16	***
X000.READ.sentence11		1.52	0.82	0.04	0.02	77.98	<2e-16	***
X000.READ.sentence12		1.05	0.74	-0.04	0.02	61.58	<2e-16	***
X000.READ.sentence13		1.26	0.78	0.00	0.02	78.52	<2e-16	***
X000.READ.sentence15		0.62	0.65	-0.13	0.02	27.55	<2e-16	***
X000.READ.sentence16		1.20	0.77	-0.01	0.03	46.42	<2e-16	***
X000.READ.sentence17		-0.97	0.28	-0.50	0.03	-36.86	<2e-16	***
X000.READ.sentence18		-0.95	0.28	-0.50	0.04	-21.90	<2e-16	***
X000.READ.sentence26	harder	-0.32	0.42	-0.36	0.03	-12.32	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

Among the indicators associated with the building of a situational model (Table 60), intentionality was not significant (X049.READ.intent), and temporality (X050.READ.temp) did not comply with theoretical expectations. The final indicator, causality (X048.READ.causal), was retained.

Table 60 Estimates for READ building a situational model indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X048.READ.causal	+ve	0.15	0.54	0.04	0.01	13.60	<2e-16	***
X049.READ.intent	+ve	-0.01	0.50	0.00	0.01	-1.38	0.17	
X050.READ.temp	+ve	-0.22	0.45	-0.05	0.00	-46.78	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

4.4.2.1.4 RD indicators

The estimates for response decision (RD) indicators are given in Table 61. All estimates were statistically significant and only one was not interpretable in terms of theory: the effect of improved performance due to use of the same relevant

text for previous items (X059.RD.pract). Indicators concerned with the semantic match between options and relevant text (X054.RD.LSA.term.KEY, X055.RD.LSA.term.DIST, X056.RD.LSA.doc.KEY, X057.RD.LSA.doc.DIST) were all significant and aligned with theoretical expectations. The indicator for dispersal of relevant information within the texts had a very small effect but it accorded with theory, and so was retained.

Table 61 Estimates for RD indicators

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X054.RD.LSA.term.KEY	+ve	0.17	0.54	0.04	0.05	3.76	0.00	***
X055.RD.LSA.term.DIST	-ve	-0.25	0.44	-0.06	0.05	-4.87	0.00	***
X056.RD.LSA.doc.KEY	+ve	0.89	0.71	0.21	0.02	35.91	<2e-16	***
X057.RD.LSA.doc.DIST	-ve	-0.08	0.48	-0.02	0.02	-3.19	0.00	**
X058.RD.disperse	-ve	0.00	0.50	0.00	0.00	-19.69	<2e-16	***
X059.RD.pract	+ve	-0.25	0.44	-0.06	0.00	-51.92	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

4.4.2.2 Summary of indicators which were retained

A summary of the indicators retained from the testing is presented in Table 62. The indicators in the table were used to determine what should be contained in the final model, described in 4.5.1.

Table 62 Indicators retained from the testing phase

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X006.OP.CELEX.cont.f	+ve	2.30	0.91	0.41	0.04	51.96	<2e-16	***
X008.OP.CELEX.cont.log	+ve	0.28	0.57	0.07	0.01	37.33	<2e-16	***
X010.OP.hypernymy	-ve	-0.05	0.49	-0.01	0.01	-5.89	0.00	***
X014.OP.mod.noun	-ve	-0.35	0.41	-0.09	0.01	-26.66	<2e-16	***
X015.OP.left.emb	-ve	-0.17	0.46	-0.04	0.01	-22.45	<2e-16	***
X018.OP.fronted1NO	easier	1.04	0.74	0.00	0.01	109.01	<2e-16	***
X018.OP.fronted2YES	harder	0.39	0.60	-0.14	0.02	22.15	<2e-16	***
X019.OP.passive	-ve	0.00	0.50	0.00	0.00	-6.63	0.00	***
X051.SEARCH.order1NO	harder	1.02	0.74	0.00	0.01	94.84	<2e-16	***
X051.SEARCH.order2YES	easier	1.10	0.75	0.02	0.01	96.68	<2e-16	***
X052.SEARCH.demarc.CLPS1LOW	harder	1.04	0.74	0.00	0.01	100.30	<2e-16	***
X052.SEARCH.demarc.CLPS2HIGH	easier	1.10	0.75	0.01	0.01	89.00	<2e-16	***
X026.READ.syll	-ve	-0.97	0.28	-0.22	0.03	-37.51	<2e-16	***
X027.READ.BNC.CLPS1LOW	easier	1.31	0.79	0.00	0.01	115.82	<2e-16	***
X027.READ.BNC.CLPS3HIGH	harder	0.92	0.72	-0.07	0.01	91.62	<2e-16	***
X028.READ.AWL0	easier	1.68	0.84	0.00	0.03	56.02	<2e-16	***
X028.READ.AWL4	harder	1.07	0.74	-0.10	0.01	109.24	<2e-16	***
X031.READ.CELEX.cont.f	+ve	0.77	0.68	0.18	0.03	27.19	<2e-16	***
X032.READ.CELEX.all.f.log	+ve	0.72	0.67	0.17	0.04	20.39	<2e-16	***
X035.READ.hypernymy	-ve	-0.09	0.48	-0.02	0.02	-6.01	0.00	***
X037.READ.lex.density	-ve	-0.02	0.50	0.00	0.00	-17.46	<2e-16	***
X043.READ.fronted.CLPS0NONE	easier	1.25	0.78	0.00	0.01	109.21	<2e-16	***
X043.READ.fronted.CLPS1LOW		0.95	0.72	-0.06	0.01	80.35	<2e-16	***
X043.READ.fronted.CLPS3HIGH	harder	0.60	0.65	-0.13	0.02	37.72	<2e-16	***
X044.READ.passive	-ve	0.00	0.50	0.00	0.00	-2.28	0.02	*
X045.READ.connect	+ve	0.00	0.50	0.00	0.00	10.03	<2e-16	***
X046.READ.stem.o	+ve	0.27	0.57	0.07	0.03	9.30	<2e-16	***
X048.READ.causal	+ve	0.15	0.54	0.04	0.01	13.60	<2e-16	***
X054.RD.LSA.term.KEY	+ve	0.17	0.54	0.04	0.05	3.76	0.00	***
X055.RD.LSA.term.DIST	-ve	-0.25	0.44	-0.06	0.05	-4.87	0.00	***
X056.RD.LSA.doc.KEY	+ve	0.89	0.71	0.21	0.02	35.91	<2e-16	***
X057.RD.LSA.doc.DIST	-ve	-0.08	0.48	-0.02	0.02	-3.19	0.00	**
X058.RD.disperse	-ve	0.00	0.50	0.00	0.00	-19.69	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

4.5 Analysis of final model

4.5.1 Model composition

As a result of the analysis specified in 3.8, a final model was specified. This model, together with the estimates and other statistics found in 4.4.2.1 is presented in Table 63. These figures represent the best estimates for each indicator, as

estimating indicators together usually implies some colinearity, which yields less accurate estimates (3.8.3.1.1). It was decided that, where indicators were essentially measuring the same thing, specifically, no more than one should be retained for each component if their impact was not easily theoretically distinguishable (3.8.2.2). For this reason, a choice had to be made between all the lexical frequency indicators which were found acceptable for OP, and, in addition, those found to be acceptable for READ. This was done simply by selecting the indicators with the largest coefficients, and, therefore, the biggest impact. For OP, this was X006.OP.CELEX.cont.f, with a figure for influence of 40.9% (Table 47); for READ, it was X031.READ.CELEX.cont.f, with a figure of 18.39% (Table 53).

As with the indicators for lexical frequency, the RD indicators concerning the match between options (key or distractors) and relevant text were also subject to a choice. In this case, it was between those indicators where the semantic match was based on individual *terms* within the text, or on the whole text (doc). As with the indicators for lexical frequency, those with larger influence were chosen.

Table 63 Indicators contained in the final model, with estimates error and significance from independent analysis (4.4.2.1)

	Expected outcome	Log odds	Probability	Influence	Std. Error	z value	Pr(>|z|)	
X006.OP.CELEX.cont.f	+ve	2.30	0.91	0.41	0.04	51.96	<2e-16	***
X010.OP.hypernymy	-ve	-0.05	0.49	-0.01	0.01	-5.89	0.00	***
X014.OP.mod.noun	-ve	-0.35	0.41	-0.09	0.01	-26.66	<2e-16	***
X015.OP.left.emb	-ve	-0.17	0.46	-0.04	0.01	-22.45	<2e-16	***
X018.OP.fronted1NO	easier	1.04	0.74	0.00	0.01	109.01	<2e-16	***
X018.OP.fronted2YES	harder	0.39	0.60	-0.14	0.02	22.15	<2e-16	***
X019.OP.passive	-ve	0.00	0.50	0.00	0.00	-6.63	0.00	***
X051.SEARCH.order1NO	harder	1.02	0.74	0.00	0.01	94.84	<2e-16	***
X051.SEARCH.order2YES	easier	1.10	0.75	0.02	0.01	96.68	<2e-16	***
X052.SEARCH.demarc.CLPS1LOW	harder	1.04	0.74	0.00	0.01	100.30	<2e-16	***
X052.SEARCH.demarc.CLPS2HIGH	easier	1.10	0.75	0.01	0.01	89.00	<2e-16	***
X026.READ.syll	-ve	-0.97	0.28	-0.22	0.03	-37.51	<2e-16	***
X031.READ.CELEX.cont.f	+ve	0.77	0.68	0.18	0.03	27.19	<2e-16	***
X035.READ.hypernymy	-ve	-0.09	0.48	-0.02	0.02	-6.01	0.00	***
X037.READ.lex.density	-ve	-0.02	0.50	0.00	0.00	-17.46	<2e-16	***
X043.READ.fronted.CLPS0NONE	easier	1.25	0.78	0.00	0.01	109.21	<2e-16	***
X043.READ.fronted.CLPS1LOW		0.95	0.72	-0.06	0.01	80.35	<2e-16	***
X043.READ.fronted.CLPS3HIGH	harder	0.60	0.65	-0.13	0.02	37.72	<2e-16	***
X044.READ.passive	-ve	0.00	0.50	0.00	0.00	-2.28	0.02	*
X045.READ.connect	+ve	0.00	0.50	0.00	0.00	10.03	<2e-16	***
X046.READ.stem.o	+ve	0.27	0.57	0.07	0.03	9.30	<2e-16	***
X048.READ.causal	+ve	0.15	0.54	0.04	0.01	13.60	<2e-16	***
X055.RD.LSA.term.DIST	-ve	-0.25	0.44	-0.06	0.05	-4.87	0.00	***
X056.RD.LSA.doc.KEY	+ve	0.89	0.71	0.21	0.02	35.91	<2e-16	***
X058.RD.disperse	-ve	0.00	0.50	0.00	0.00	-19.69	<2e-16	***

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

4.5.2 Examination of model assumptions

4.5.2.1 Adequate dimensionality

The scree plot for this model, shown in Figure 10 is very close to that for the equivalent Rasch model (Figure 9). For the same reasons given for that model (4.4.1.1), this model was taken to adequately account for the dimensional structure of the data.

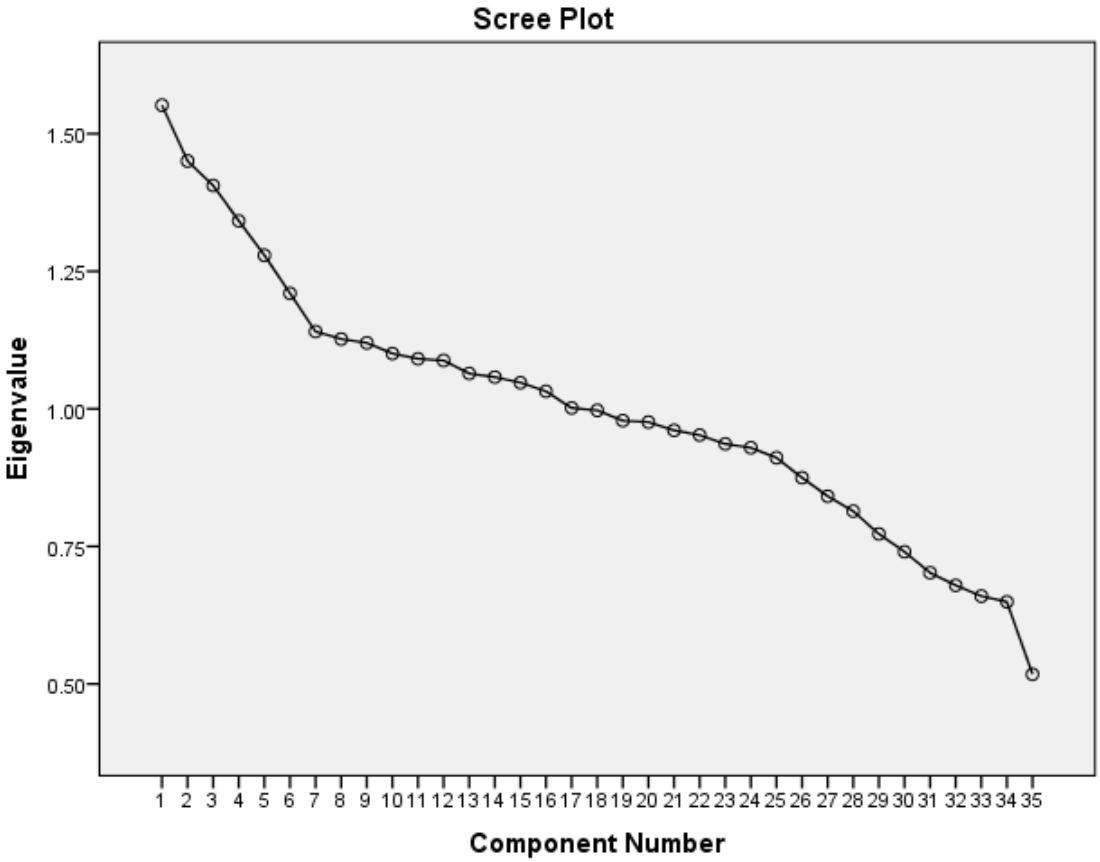


Figure 10 Scree plot, final LLTM with four dimensions and correction for dependency between items 17 and 18

4.5.2.2 LD

The results of the investigation of violations of LD are given in Table 64, Table 65, Table 66 and Table 67. The pairing with the largest Q_3^2 index was 1 and 3 (Table 64), followed by that between 16 and 20, with 8% each of shared variance, respectively. The dependency between items 17 and 18, still identified as the largest in the corresponding Rasch model (4.4.1.1), was, in the LLTM model, 0. It was assumed that the covariance between the two items was somehow accounted for by other fixed effects. In all, the level of LD was considered low and no further corrections were added to the model.

Table 64 Q32 index, final LLTM with four dimensions and correction for dependency between items 17 and 18, Part 1

	X1	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*
X1	1.00						
X2	0.02	1.00					
X3	0.08	0.02	1.00				
X4	0.01	0.01	0.00	1.00			
X5	0.01	0.00	0.00	0.01	1.00		
X6	0.01	0.01	0.00	0.00	0.03	1.00	
X7	0.01	0.01	0.00	0.00	0.00	0.04	1.00

Table 65 Q_3^2 index, final LLTM with four dimensions and correction for dependency between items 17 and 18, Part 2

	X8	*X9*	*X10*	*X11*	*X12*	*X13*	*X14*	*X15*
X8	1.00							
X9	0.00	1.00						
X10	0.00	0.00	1.00					
X11	0.00	0.00	0.00	1.00				
X12	0.00	0.00	0.00	0.00	1.00			
X13	0.00	0.00	0.00	0.00	0.00	1.00		
X14	0.00	0.00	0.00	0.00	0.00	0.00	1.00	
X15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 66 Q32 index, final LLTM with four dimensions and correction for dependency between items 17 and 18, Part 3

	X16	***X17***	***X18***	***X19***	***X20***	***X21***	***X22***
X16	1.00						
X17	0.00	1.00					
X18	0.00	0.00	1.00				
X19	0.00	0.00	0.01	1.00			
X20	0.08	0.00	0.01	0.01	1.00		
X21	0.01	0.01	0.01	0.00	0.01	1.00	
X22	0.00	0.00	0.00	0.01	0.01	0.04	1.00

Table 67 Q_3^2 index, final LLTM with four dimensions and correction for dependency between items 17 and 18, Part 4

	X23	*X24*	*X25*	*X26*	*X27*	*X28*	*X29*	*X30*	*X31*	*X32*	*X33*	*X34*	*X35*
X23	1.00												
X24	0.00	1.00											
X25	0.02	0.00	1.00										
X26	0.00	0.02	0.00	1.00									
X27	0.00	0.00	0.00	0.00	1.00								
X28	0.00	0.00	0.00	0.02	0.00	1.00							
X29	0.00	0.00	0.01	0.00	0.01	0.00	1.00						
X30	0.01	0.00	0.00	0.00	0.00	0.00	0.00	1.00					
X31	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	1.00				
X32	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	1.00			
X33	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	1.00		
X34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	1.00	
X35	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

4.5.2.3 *Normality of the distribution of candidate ability estimates*

Descriptive statistics for the distribution of candidate ability estimates are available in Table 68. As with corresponding statistics in Table 46, they show that the distribution is approximately normal. In the case of all dimensions, the mean and median are close to each other and close to 0. Furthermore, skew and kurtosis are both within a range of +/2. This assumption was therefore considered to have held.

Table 68 Summary statistics for candidate ability estimates for each dimension of the final LLTM with four dimensions and correction of dependency between items 17 and 18

	Part 1	**Part 2**	**Part 3**	**Part 4**
Mean	-0.02	-0.04	-0.04	-0.03
Median	0.00	-0.05	-0.13	-0.04
SD	0.96	0.73	1.06	0.70
Variance	0.92	0.54	1.11	0.49
Max	1.95	1.67	2.22	1.61
Min	-3.24	-2.30	-3.12	-2.52
Range	5.19	3.97	5.34	4.14
Skew	-0.23	0.03	0.12	0.00
Kurtosis	-0.35	-0.41	-0.69	-0.38

It is additionally worth noting that Parts 1 and 3 have the largest standard deviations. One reasons for this is likely to be because these parts exhibit some degree of dependency between items. According to University of Cambridge ESOL Examinations (2008), when LD is violated, the standard deviation of ability estimates increases. This is intuitively reasonable, as the probability of obtaining the same score on two dependent items is larger. This causes the measurement scale to stretch, as more candidates get either a score of 1 on both items, or a score of 0.

4.5.2.4 *Absence of colinearity between the fixed effects*

Absence of colinearity between fixed effects is perhaps the least important assumption, as, according to De Boeck et al. (2011), it only affects the coefficients for the fixed effects, and not the more global statistics of the model. The latter are

of interest here, as values for coefficients were taken from the models estimated for individual indicators 4.4.2.1, for among other reasons to avoid the possibility of colinearity (3.8.2.2). The cut off value for excessive colinearity was set at +/- 0.866 (3.8.3.1.1). Correlation coefficients are shown in Table 69 and Table 70. The names of indicators across the top and down the left identify cells containing the value for particular combinations of indicators. In these tables, no values exceed 0.63, or are lower than -0.78. The latter constitutes 61.4% shared variance but is well below the cut off value and therefore does not represent a significant violation of the assumption. Overall, the colinearity represents relatively minor overlap between indicators, with a mean of 5.8% shared variance and a standard deviation of 8.6%.

Table 69 Correlations between fixed effects, final LLTM model, first 10 indicators

	X006.OP.CELEX.cont.f	X010.OP.hypernymy	X014.OP.mod.noun	X015.OP.left.emb	X018.OP.fronted1NO	X018.OP.fronted2YES	X019.OP.passive	X051.SEARCH.order2YES	X052.SEARCH.demarc.CLPS2HIGH	X026.READ.syll	X031.READ.CELEX.cont.f	X035.READ.hypernymy
X006.OP.CELEX.cont.f	1											
X010.OP.hypernymy	-0.2	1										
X014.OP.mod.noun	0.34	-0	1									
X015.OP.left.emb	-0.6	0.06	-0.1	1								
X018.OP.fronted1NO	-0.3	-0.3	-0.2	0.45	1							
X018.OP.fronted2YES	0.48	-0.2	-0	-0.8	0.03	1						
X019.OP.passive	-0.3	-0.3	0.22	0.6	0.39	-0.3	1					
X051.SEARCH.order2YES	-0.2	0.32	0.04	0.22	-0.6	-0.6	-0	1				
X052.SEARCH.demarc.CLPS2HIGH	0.44	0.16	0.42	-0.5	-0.3	0.41	-0.4	-0.3	1			
X026.READ.syll	-0.1	-0	-0.4	-0.1	-0.3	0.01	-0.2	0.32	-0.1	1		
X031.READ.CELEX.cont.f	-0.1	-0.2	-0.3	-0.1	0.01	0.06	-0.2	0.09	-0.1	0.4	1	
X035.READ.hypernymy	0.37	0.16	0.17	-0.4	-0.3	0.22	-0.1	-0	0.37	-0.1	0.01	1

Table 70 Correlations between fixed effects, final LLTM model, last 9 indicators and correction for LD (dep51)

	X006.OP.CELEX.cont.f	X010.OP.hypernymy	X014.OP.mod.noun	X015.OP.left.emb	X018.OP.fronted1NO	X018.OP.fronted2YES	X019.OP.passive	X051.SEARCH.order2YES	X052.SEARCH.demarc.CLPS2HIG...	X026.READ.syll	X031.READ.CELEX.cont.f	X035.READ.hypernymy	X037.READ.lex.density	X043.READ.fronted.CLPS1LOW	X043.READ.fronted.CLPS3HIGH	X044.READ.passive	X045.READ.connect	X046.READ.stem.o	X048.READ.causal	X055.RD.LSA.term.DIST	X056.RD.LSA.doc.KEY	X058.RD.disperse	dep51
X037.READ.lex.density	-0.3	0.04	0.16	0.25	-0.2	-0.1	0.53	0.17	-0.1	-0.1	-0.2	-0	1										
X043.READ.fronted.CLPS1LOW	0.2	0.1	0	-0.1	-0.5	-0	-0.1	0.17	0.21	0.12	-0.3	0.13	0.21	1									
X043.READ.fronted.CLPS3HIGH	-0	0.01	-0.1	-0.1	-0.1	0.04	0.03	0.03	-0.3	-0.1	-0.2	0	0.15	0.12	1								
X044.READ.passive	-0.4	-0.3	-0.1	0.14	0.1	-0.1	0.27	0.13	-0.4	0.2	0.37	-0.3	0.36	-0.3	0.15	1							
X045.READ.connect	-0.2	-0	-0.5	-0	-0.2	-0	-0.2	0.29	-0.4	0.34	0.39	0.12	0.06	0	0.35	0.41	1						
X046.READ.stem.o	-0.3	0.21	0	0.06	-0.3	-0.2	0.02	0.28	0.07	0.13	0.04	-0	0.34	0.42	0.03	0.07	-0	1					
X048.READ.causal	0.2	0.13	0.4	-0.3	-0.5	0.13	0.05	0.14	0.45	0.14	-0.2	0.4	0.46	0.39	0.08	0.02	-0.1	0.36	1				
X055.RD.LSA.term.DIST	-0.2	0.52	0.16	0.2	-0.1	-0.1	0.3	0	0.16	-0	-0.3	0.16	0.54	0.28	0.05	-0.1	-0.1	0.34	0.4	1			
X056.RD.LSA.doc.KEY	-0.1	-0.1	-0.1	0.3	0.22	-0.1	0.39	-0.1	-0.3	-0.2	-0.1	0.06	0.19	0.06	0.05	-0	0.01	-0.1	-0.1	-0.2	1		
X058.RD.disperse	-0.3	-0.3	0.1	0.29	-0	-0.1	0.57	0.22	-0.4	0.08	0.15	-0.3	0.63	-0.1	0.07	0.41	0.07	0.12	-0.1	0.12	0.3	1	
dep51	0.08	-0	-0	-0.1	0.1	0.03	-0.1	-0.1	0.06	-0.1	0.21	0.06	-0.1	-0.2	-0	0.01	0.06	-0.3	-0.1	-0.1	-0.1	-0.1	1

4.5.3 Results for subcomponents and components

As discussed in 3.8.2.2, values estimated in the individual analyses of the indicators do not suffer from colinearity, and were therefore used in the assessment of the influence of components, subcomponents and attributes in the test, rather than the values estimated with the final LLTM itself. Figures for influence are collated by subcomponent and component, as set out in 2.6) in Table 71 and Table 72 respectively. The component is given in the first column of each table, with the subcomponents listed next in Table 71. Further columns provide the pooled influence for positive and negative indicators within the component and subcomponents and the net effect is also shown. The figures for subcomponent show that no evidence of word recognition was found. Lexical access for the OP component was the most influential of all subcomponents, having an influence of around twice that of the next largest (READ: word recognition).

Table 71 Collation of the influence of fixed effects by subcomponent

Component	Subcomponent	Influence		
		+ve	-ve	Net influence
OP	Word recognition	0.00%	0.00%	0.00%
	Lexical access	40.90%	-1.17%	39.73%
	Syntactic parsing	0.00%	-12.89%	-12.89%
	Establishing propositional meaning	0.00%	-14.27%	-14.27%
SEARCH	LSA match	0.00%	0.00%	0.00%
	Item order	1.53%	0.00%	1.53%
	Demarcatedness	1.20%	0.00%	1.20%
READ	Word recognition	0.00%	-22.44%	-22.44%
	Lexical access	18.39%	-2.73%	15.66%
	Syntactic parsing	0.00%	0.00%	0.00%
	Establishing propositional meaning	0.00%	-13.16%	-13.16%
	Establishing a coherent textbase	6.62%	0.00%	6.62%
	Building a situational model	3.68%	0.00%	3.68%
RD	Option match	20.86%	-6.14%	14.72%
	Dispersal	0.00%	-0.06%	-0.06%
	Practice	0.00%	0.00%	0.00%

Some indicators were not identified for all subcomponents, and OP: word recognition, SEARCH: LSA match, READ: syntactic parsing and RD: practice were all left out.

Based on the figures in Table 71, the absolute influence of subcomponents in the READ component is represented graphically in Figure 11. This is of interest because it may be understood as an indication of the relative influence of each

subcomponent, something of interest to other researchers and discussed in 2.2.2. The figure shows that the evidence found in the current study indicates greater influence for lower level processes than for higher level processes. Since, as shown in 0, not all variance is explained by the current study, the results displayed in the figure cannot be understood as a comprehensive finding, but rather as an indication. Missing variance may explain syntactic parsing, and may also alter the rank of some of the subcomponents investigated.

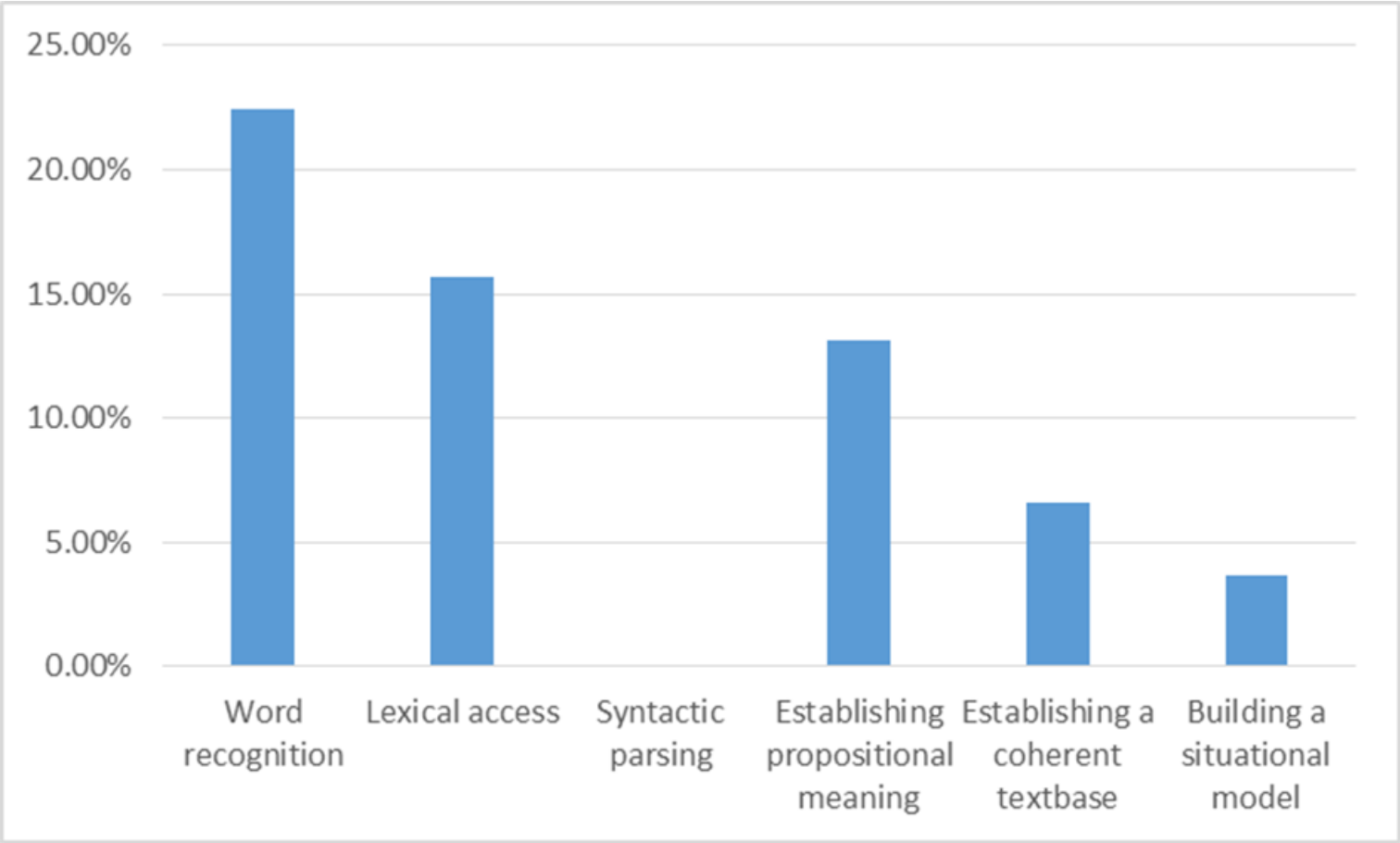


Figure 11 Influence (absolute) of subcomponents in READ

Table 72 Collation of the influence of fixed effects by component

	Influence		
Component	**+ve**	**-ve**	**Net influence**
OP	40.90%	-28.34%	12.56%
SEARCH	2.73%	0.00%	2.73%
READ	28.70%	-38.34%	-9.64%
RD	20.86%	-6.20%	14.66%
Net	93.19%	-72.88%	20.31%

Overall, as shown in Table 72, the indicators retained in the final model represented attributes which increased the chance of success on items. The SEARCH component had a particularly small influence, compared to the other components.

4.5.4 Variance explained

4.5.4.1 Items and fixed effects

As described in 3.8.3.1.3, the variance explained by the item difficulty portion of the final LLTM was assessed in relation to its lowest possible limit (an empty model) and its highest (an equivalent Rasch model). For all three models, the random effects and corrections for violations of LD were retained. Each model was compared using the LRT, to determine whether there was a significant difference in fit. The percentage of variance was then explained by subtracting the deviance of the final LLTM from the reference model and dividing the result by the difference between the deviance of the empty model and the Rasch model. This has the consequence of placing the variance explained by the final LLTM on a scale from the variance explained by the empty model to that explained by the Rasch model.

The results of the LRTs are given in Table 73 and Table 74. They show that, as expected, the LLTM fits better than the empty model, and the Rasch model fits better than the LLTM. The deviance of the all three models is also displayed in these tables. The results of applying the formula described above to the model deviances are given in the final column of each table (variance explained). This shows it that the proportion of variance explained by the fixed effect attributes of

the LLTM was 75.79% of the variance explained by the fixed effects of the Rasch model (3.8.3.1.3).

Table 73 LRT of two models with four dimensions and correction for dependency between items 17 and 18: the empty model and the final LLTM

	Model statistics					LRT				Variance explained
	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)		
Final.4d.dep5	33	351802	352156	-175868	351736					
Rasch.4d.dep5	46	342802	343295	-171355	342710	23383	14	<2.2e-16	***	24.21%

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

Table 74 LRT of two models with four dimensions and correction for dependency between items 17 and 18: the final LLTM and the Rasch model

	Model statistics					LRT				Variance explained
	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)		
Empty model 4d.dep5	12	380009	380138	-189993	379985					
Final.4d.dep5	33	351802	352156	-175868	351736	28249	21	<2.2e-16	***	75.79%

Signif. codes: '***'=0.001 '**'=0.01 '*'=0.05 '.'=0.1 ' '=1

4.6 Chapter summary

In this chapter, results pertaining to item attributes were given (2.7). The results were also shown to relate to subcomponents and components specified in the theoretical model (2.6). Results which were significant and interpretable in terms of the theoretical model provide a foundation for understanding the construct representation of FCE Dec 05, and this will be elucidated further in 5. An attempt to quantify the amount of variance explained by the model is also of interest, as this implies two things:

- the amount of what is currently unknown about the test and must be explained by future studies
- the utility of the method of investigation in the current study

Both will be addressed in Chapter 5.

5 Discussion and Conclusions

5.1 Introduction

In this chapter, conclusions will be drawn concerning the results of the study (see Chapter 4), limitations of the methodology and implications of the findings will be outlined and suggestions made for future research.

The study investigated the construct of a Reading test (FCE) by analysing one form of that test (December 2005) in relation to a composite theoretical model of reading in a second language. It involved an assessment of the relationship between attributes (described in 3.6.1.8) related to the test materials, such as lexical density or demarcation of text in the reading passage, (Appendix 1: test papers) and the difficulty of items. Theoretical models, described in 2.2.1, 2.3 and 2.4, were used as a basis for determining which attributes should be included in the study and how they were expected to behave (2.7). The attributes were grouped according to subcomponents and components suggested by the theoretical models. These are operationalised stages in the complex cognitive process of responding to items and based on a model by (Rouet, 2012). The first stage is termed *OP* ('options') in the current study, and involves the formation of a task model, or the setting of goals and the determination of the means with which to achieve them. This is expected to be based, in large part, on reading of the text in the stem of the item and in the options. For this reason, this text was used to formulate the attributes for this component. The model of reading applied was based on the Khalifa and Weir (2009) model, which defines a number of subcomponents in which attributes nest 2.6. The next component is *SEARCH*, where candidates use their newly-formed task model to locate relevant text to respond to the item. After this, candidates read the text they have located as potentially relevant. This component is termed *READ* and is based on the same model as OP, except that there are more subcomponents and attributes because

the text is usually longer and requires additional processing steps. Finally, as suggested by Embretson and Wetzel (1987) as well as Rouet (2012), a decision over which response to select is required and this is referred to as response decision, or RD.

In order to measure the impact of attributes on item difficulty, indicators (variables) are constructed to be analysed quantitatively using a Linear Logistic Test Model (LLTM) (see 2.9.3.3.3). This model is a variant of the Rasch model, but, instead of specifying a single difficulty parameter for the item, a parameter for each indicator is included. The difficulty parameter is effectively decomposed into facets of difficulty. The result is that the contribution of each attribute to item difficulty is estimated by the model.

The appraisal of the results will centre on the research questions introduced in 2.10 and included again here:

1. Which contextual attributes (see 2.7) can be shown to influence the difficulty of FCE Dec 2005, and by how much?
2. Which subcomponents included in the composite model (see 2.7) can be shown to influence the difficulty of FCE Dec 2005, and by how much?
3. Which components included in the composite model (see 2.7) can be shown to influence the difficulty of FCE Dec 2005, and by how much?
4. What evidence can be found of test methods effects influencing item difficulty?
5. What proportion of the variance of the corresponding Rasch model does the LLTM account for?

The results, and therefore the conclusions, for questions 1 to 3 for any given component are highly interrelated because indicators are nested in subcomponents which are, in turn, nested in components. To focus, for example, on the subcomponent of lexical access for the READ component (2.7.1.2) requires reference to the results for the indicators. Discussion at the component level will necessarily refer to the subcomponents and indicators. For this reason,

conclusions are also arranged in a nested fashion, with questions 1 to 3 being addressed for a single component, its subcomponents and indicators in a single section. Consideration of the other questions follows, as does a discussion of limitations and suggestions for further research.

5.2 Research questions 1, 2 and 3: indicators, subcomponents and components

For these research questions, comments will be grouped according to the components and subcomponents they relate to. This is because indicators, subcomponents and components form a nested structure, and what can be said about the significance of indicators within a subcomponent or component also has relevance when discussing them. The results for indicators and components will therefore be discussed together. The impact on components will be considered after this.

5.2.1 OP

5.2.1.1 OP indicators and subcomponents

For the OP component, indicators covering the subcomponents *word recognition*, *lexical access*, *syntactic parsing* and *establishing propositional meaning* were tested. Word recognition is the initial process of identifying the word through its graphical form and lexical access is the retrieving of the word from the mental lexicon. After this, the syntactic structure of the group of words, or proposition, is decoded and the grammatical function of words identified (syntactic parsing). Finally, the information gained so far is combined to determine a small unit of semantic meaning (establishing propositional meaning). Each attribute is described in more detail in 2.7.1.

The results for the OP component are summarised in 4.4.2.1.1. They show that seven indicators were found to affect item difficulty in an interpretable way. Six of these indicators were retained for the final model, and are listed in Table 47, Table 48, Table 49 and

Table 50. The Seventh was omitted from the final model for reasons given in 4.4.2.1.1.

For each subcomponent, results are summarised in Table 71. For the OP component, of the subcomponents tested, lexical access, syntactic parsing and establishing propositional meaning were represented in the final model.

In the remainder of this section, the characteristics of the each indicator and nature of the modelled subcomponent are discussed.

5.2.1.1.1 Word recognition

The first indicator, X001.OP.syll, or the mean number of syllables per word, was not retained for the final model because the direction of influence was not as expected. As described in 2.7.1.1, a higher number of syllables is expected to be associated with increasing difficulty because they provide an important clue to word recognition. Furthermore, an increase in processing times has been observed to coincide with an increase in the number of syllables (Weir, 2013).

Scrutiny of Appendix 11: incidence matrix summary, however, shows that, on average, items in Part 4 of the test have words with more syllables (1.75) than the other parts (1.50). This is perhaps because phrases such as

doing considerable background research

are commonly found in Part 4. These use techniques such as nominalisation to summarise meaning. This often has the effect of producing sentences made up of fewer, but longer words. Part 1 contains the next highest average number of syllables (1.57), and again the task aims to summarise large sections of text in brief statements. It may also be that the words with more syllables were relatively common and instantly recognisable, thereby mitigating the need to decompose a word into syllables in order to recognise it. In these cases, item difficulty would not be expected to correlate with the number of syllables.

5.2.1.1.2 Lexical access

Several indicators related to lexical access were tested and are described more fully in 2.7.1.2. Of the indicators concerning word frequency, only indicators derived from CELEX (a 17.9 million word data base of word frequencies) were found to be suitable for inclusion in the final model. This may be due in part to the way in which the indices were constructed. The other sources of information were the British National Corpus (BNC), a corpus of 100 million words, and the Academic Word List (AWL), with frequency information on 2,570 words which have been found to occur frequently in academic texts. Indices were based on frequency tiers, such that, in the case of the BNC, words were grouped according to their membership of the most frequent 1,000 words in the corpus or of the second, or third or other most frequent 1,000 words (2.7.1.2). For the AWL indicator, the first two tiers were with the same as those of the BNC minus any words found in the AWL itself. The third tier comprised words found in the AWL. An alternative approach would have been to give each word in the corpus a measure of frequency per 1,000 words, in a similar way to the indices of CELEX. The frequency tier approach results in a less precise measure of frequency, but in some contexts this would be an advantage. For example, words which are reasonably close in frequency may not have a significantly different impact on reading difficulty, so an index which differentiates them would offer only false precision. Based on the analysis, however, the CELEX approach provided a measure of frequency which more effectively captured the effect of lexical frequency on item difficulty.

An additional, interrelated reason the CELEX-based frequency indices functioned better than those of the BNC and AWL could be that the indicators for the latter two were factors, rather than continuous variables (see 3.7.1 for a discussion of the difference). Scrutiny of Appendix 11: incidence matrix summary shows that the BNC and AWL indicators varied between the test parts. This suggests that the coefficients were influenced by the nature of the words related to each task. As the options related to Parts 1 and 3 were the same for all items in these parts (all the options were treated as equally applicable to each item – see 3.6), it is likely

that a factor indicator was simply too crude to reflect the impact of word frequency on item difficulty for these parts. Compared to the continuous CELEX-based indicators, less information is available in the indicator.

Three indicators derived from the CELEX database were tested. Among them, the coefficient for the log frequency of all words (X006.OP.CELEX.all.f.log) was not considered interpretable, whereas those concerning content words were. The difficulty with the indicator for the log of all words may be understandable in relation to the coefficient for lexical density (X012.OP.lex.density), which was not accepted as it was positively associated with item difficulty. Examination of Appendix 11: incidence matrix summary, reveals much larger values for the OP text in relation to those for the READ text. These are no doubt due in part to the lack the requirement for complete sentences among the options and item stems, which led to function words being omitted in Parts 1, 2 and 4. By omitting function words, the understanding of content words becomes that much more important, as the support for inferring information about words given by the grammatical structure of the sentence, such as part of speech or semantic meaning, is lessened. In this way, the indicator for content word frequency (X006.OP.CELEX.cont.f) became the most important indicator in this subcomponent.

The remaining three indicators related to the semantic characteristics of words and are self-explanatory: polysemy (X011.OP.polysemy), hypernymy (X010.OP.hypernymy) and concreteness (X013.OP.concrete). Of these three, only the indicator for hypernymy was retained, although it had only a very slight influence (-1.17%). As discussed in 2.7.1.2, some hypernyms are relatively more frequent, and therefore easier to process, than their subordinates, leading to a higher index increasing item difficulty. Other hypernyms are relatively less frequent than their subordinates, resulting in a lower index increasing item difficulty. These results suggest that the former tendency predominated in the items examined in the current study.

Evidence of large variance between tasks for polysemy and concreteness was evident in Appendix 11: incidence matrix summary. One explanation for lack of influence on item difficulty is that these indicators did not differentiate greatly between candidates in the sample. Variation of the indicators between test parts is likely to be random noise due to the small size of data for these texts. In addition, in the case of concreteness, as discussed in 2.7.1.2, the nature of the word list upon which the indices are based is somewhat suspect, and error in the word list may also therefore contribute to the coefficients obtained for this index.

In sum, in the final model, measured lexical access consists of the frequency of content words. There is some evidence to suggest that content words are particularly important because some function words are omitted from the texts of the item stem and of the option (for example, in Parts 1, 2 and 4).

5.2.1.1.3 Syntactic parsing

Both indicators associated with syntactic parsing, noun modification (X014.OP.mod.noun) and left embeddedness (X015.OP.left.emb), were interpretable for the OP text. A larger number of words modifying the noun is expected to increase difficulty, just as more words before the main verb (left-embeddedness) are. Both are described more fully in 2.7.1.3. Among the text for OP, that for Part 4 had the highest indicator values for modified noun phrases (see Appendix 11: incidence matrix summary), for example, item 28:

doing considerable background research

included two adjectives modifying the noun. High modification would appear to be a task feature, as good distraction will have some features associated with the stem but only the key will provide a complete match. 'Background research,' it could be argued, is at least implied for all options of item 28 (see Appendix 1: test papers), but is only clearly 'considerable' in the case of A, the key.

Part 3 has the highest level of left embeddedness according to Appendix 11: incidence matrix summary. Option B, in particular, included a large number of words (22) before the main verb phrase (*was getting noticed*):

> *But his interest in this, the world's most widespread kingfisher and the only member of its cosmopolitan family to breed in Europe, was getting noticed.*

Although option B is the most extreme case, other options suggest that left embeddedness may be a task feature. Khalifa and Weir (2009:72) state, these items aim to determine that candidates can 'understand how examples are introduced and changes of direction signalled'. At such points, contrastive connectors, such as in option B ('But'), explicit references, such as that in option A ('This')

> *This is why a kingfisher may appear…*

or a marked, fronted structure, such as that in option C

> *A sure sign of his depth of feeling for this little bird is…*

are more likely to be present before the main verb, thus increasing left embeddedness.

The subcomponent may be summarised as being important in particular tasks, due to features of these tasks.

5.2.1.1.4 Establishing propositional meaning

Only two of the indicators associated with establishing propositional meaning (2.4.1.1.4) yielded interpretable values for the OP text. The number of propositions (X022.OP.props) was a factor-based indicator with 13 levels. It may be, therefore, that this indicator did not provide a suitable index for the reasons given in 5.2.1.1.2. Propositional density (X000.OP.props.dens), or the number of propositions per word, is also uninterpretable. As it is also based on a count of propositions, this may be attributable to the same issue.

Instances of negation (X016.OP.neg), or holistic negation (X017.OP.hol.neg) and frontedness (X018.OP.fronted) were rare. Negation and holistic negation are distinguished by the former concerning only grammatically-based formulations, whereas the latter included semantic ones such as 'deficit'. Fronted structures are those where non-standard word order brings elements to the front of a sentence, whereas they would normally have a later position (e.g. cleft sentences). Only the indicator for frontedness produced interpretable results and was retained.

The indicator for passive voice (X019.OP.passive) was retained for the final model despite having a very small coefficient. As noted in 2.7.1.4, instances of passive voice may increase difficulty in a number of subcomponents, not only that of establishing propositional meaning.

5.2.1.2 The OP component

The results for OP were characterised by the nature of the test tasks. Lexical access and syntactic parsing were important, probably because of the nature of the option and stem text for some tasks. The importance of task artefacts in determining the contribution of the OP text to difficulty is not altogether surprising. As discussed in 2.3, if the function of the OP text is to allow candidates to form a task model, it should not be so difficult as to hinder their progress onto the reading of the main passage, which is surely the intended focus for the construct of reading ability. The results for the OP component show that this is the case for most indicators, although task-specific effects do affect difficulty for syntactic parsing and lexical access.

5.2.2 SEARCH

5.2.2.1 SEARCH indicators

For this component, three indicators were tested. The first involved the semantic links, established through Latent Semantic Analysis (LSA), between each stem and option text and the relevant text in the reading passage (X053.SEARCH.LSA.term). LSA is an approach to analysing text which only semantic relationships within and between texts are considered, rather than other important features, such as grammar. The other indicators were the correspondence in order of items and

relevant text (X051.SEARCH.order), and the demarcation of the relevant text (X052.SEARCH.demarc). These indicators were not divided into subcomponents, however, as there was no practical or theoretical rationale for this.

Of the indicators for this component, search order (X051.SEARCH.order) and demarcatedness (X052.SEARCH.demarc.CLPS) were found to be interpretable (4.4.2.1.2), although coefficients were very weak. In other words, order and demarcation made the items easier but only slightly. The indicator relying on semantic links between options and relevant text (X053.SEARCH.LSA.term), was not retained and therefore, did not corroborate Freedle and Kostin's (1993) findings regarding the importance of this semantic link.

5.2.2.2 The SEARCH component

For this component, two indicators were significant and interpretable (4.4.2.1.2): search order (X051.SEARCH.order), which dealt with the correspondence between items and relevant text, and demarcatedness of the reading passage (X052.SEARCH.demarc.CLPS). In both cases, influence was minor, however, it is worth noting that each effect is closely linked to test method. In the former case, only Part 4 does not exhibit a correspondence with item order; for the latter, only Part 1 and two items in Part 2 are demarcated. This effect need not be considered construct-irrelevant, however, as, in both cases, real-life scenarios could be imagined where the order of information, or demarcatedness is present or absent. For example, specific genres such as obituaries often follow an approximately chronological order, other articles do not, and so the information required may not be in a predictable order. As far as demarcated text is concerned, textbooks might contain sections separated from the flow of text, containing a specific type of information.

5.2.3 READ

5.2.3.1 READ indicators and components

Indicators for five subcomponents were tested: word recognition, lexical access, syntactic parsing, establishing propositional meaning, *establishing a coherent textbase* and *building a situational model*. As described more fully in 2.7.1.5 and

2.7.1.6, respectively, the latter two involve combining propositions into a larger structure (textbase) and, with the textbase, constructing a mental understanding of the situation described by the text, independent of the words of the text (situational model). Of these subcomponents, only syntactic parsing did not furnish indicators for the final model. Twelve indicators were found to influence item difficulty in an interpretable way, and nine are contained in Table 63. The remaining indicators were not included in the final model for reasons discussed in 4.5.1. The extent to which each subcomponent influenced item difficulty is summarised in Table 71.

5.2.3.1.1 Word recognition

The indicator for the mean number of syllables per word for the READ component (X026.READ.syll), in contrast to that for the OP component, was found to have a significant, negative relationship to item difficulty, and was therefore retained. The reasons why this indicator was not retained for OP were given in 5.2.1.1.1. In READ, however, the relationship between the number of syllables and item difficulty is strong, perhaps for related reasons. Specifically, in the reading passage, unlike for OP, words with more syllables may not be frequent and instantly recognisable. As a result, the words must be decomposed into syllables to be recognised, so word length (in syllables) did have an impact on cognitive processing.

5.2.3.1.2 Lexical access

For the READ text, the log of the frequency of the content words (X033.READ.CELEX.cont.log) was the only CELEX indicator which was not interpretable. The indicator for the log of the frequency of all words (X032.READ.CELEX.all.f.log) was the most important indicator for this subcomponent, in contrast with that for the OP component, where it was the frequency ofcontent words. This difference suggests that content words were far more important than function words.

The hypernymy and polysemy coefficients for the READ text show a similar pattern to those of the OP text. Hypernymy (X035.READ.hypernymy) had a small

coefficient, whereas Polysemy (X036.READ.polysemy) had a positive relationship with item facility, and was therefore not interpretable in terms of prevailing theory. As with the corresponding indicators for the OP text, these features simply did not distinguish between candidates in this context. The concreteness coefficient for the READ text (X038.READ.concrete) appears to be a similar case, where, although it did have a significant and interpretable coefficient, it was small and so did not distinguish between candidates well.

In sum, as with the OP component, measures of word frequency were found to be most important for this subcomponent. Other indicators were of minimal impact.

5.2.3.1.3 Syntactic parsing

Neither indicator for syntactic parsing was interpretable for the READ text, showing relative lack of influence over difficulty. This contrasts with the OP text, where parsing difficulty clearly affected item difficulty. In order to explain this, it is suggested that the reasons for their importance for OP are examined (5.2.1.1.3): test method effects distorted the linguistic features of the text in OP, which in turn made syntactic parsing harder. The results for syntactic parsing in the READ component also seem to suggest that the standard syntax found in the reading passage was not a challenge for candidates around B2 level.

5.2.3.1.4 Establishing propositional meaning

For the READ component, the indicator for fronted sentences was found to be interpretable after some of its categories were collapsed (X043.READ.fronted.CLPS). This resulted in three categories: no frontedness, one or two instances, and three or four instances. These findings match Freedle and Kostin's (1993) findings, discussed in 2.7.1.4, that frontedness increased item difficulty in their data.

The other indicators for establishing propositional meaning for the READ text were found not to be interpretable. For the number of propositions (X047.READ.props), there is some evidence of higher counts for Parts 1 and 3 (Appendix 11: incidence matrix summary). Nevertheless, unlike with the indicator for fronted structures,

this pattern did not facilitate the collapsing of categories and the recovery of coefficients with a monotonic pattern of increase. The other indices, propositional density (X000.READ.prop.dens), negation (X041.READ.neg) and holistic negation (X042.READ.hol.neg), were not interpretable and it is concluded, therefore, the presence of these attributes did not affect the difficulty of items in important ways. The indicator for passive voice (X044.READ.passive) was retained, as it was found to be interpretable and significant, though with a minimal influence (-0.04).

As discussed in 2.7.1.4, it may also be interesting to examine the frontedness and passive voice indicators with those related to syntactic parsing. All increase syntactic complexity but are categorised separately in the current study according to the stage of cognitive processing where their influence is expected to be felt most. The two indicators which were found to produce interpretable results for the READ component (X043.READ.fronted.CLPS and X044.READ.passive) are also associated with non-standard word order. It may be, therefore, that non-standard word order could be a criterial feature at B2 level, following Hawkins and Buttery's (2012) identification of cleft sentences in productive texts as criterial. Such a hypothesis would explain why syntactic features with standard word order (5.2.3.1.3) appeared to have little effect.

5.2.3.1.5 Establishing a coherent textbase

Only the READ text was examined for a relationship between the difficulty of establishing a coherent textbase and item difficulty. The two indicators dealing with connections between parts of the text, the incidence of explicit connective devices (X045.READ.connect) and noun to lemma overlap between sentences (X045.READ.stem.o), were found to have coefficients which accorded to expectations, although the coefficient for the former was very small. This contrast suggests that connectives did not discriminate well between strong and weak candidates for this test, whereas semantic cohesion did.

Both the indicator representing the type-token ratio (X034.READ.type.tok) and that for the count of sentences (X000.READ.sentence) were not interpretable. The

problematic nature of the former, that short texts result in quite unstable figures, is noted in 2.7.1.5. This may be part of the reason the type-token ratio was found not to influence difficulty greatly.

The count of sentences (X000.READ.sentence) closely resembles that of propositions (X047.READ.props) as the frequency of each is likely to correlate highly. As a result, it is not surprising that X000.READ.sentence is not interpretable, given that X047.READ.props is not. The number of sentences also has a special significance within the Khalifa and Weir (2009) model, as it is used to distinguish between local and global reading. As the indicator counts sentences within the relevant text for each item, it can be concluded that, although global reading by this definition is required, it does not increase the difficulty of items in the test. It should also be remembered, however, that when relevant text was selected, only whole sentences were included (see 3.5.1). In order to investigate further, a more nuanced approach may be required.

In sum, semantic cohesion was the most important attribute found for this subcomponent.

5.2.3.1.6 Building a situational model

Three indicators were specified for this subcomponent: the situation dimensions of temporality (X050.READ.temp), causality (X048.READ.causal) and intentionality (X049.READ.intent). As the text is read, the reader follows one or more of such dimensions and disjunctions in any dimension would be expected to increase difficulty (see 2.7.1.6). The results show that only causality was interpretable, and therefore suggests that the other dimensions did not discriminate between candidates. It is not clear why this might be, but may be related to features of the specific reading passages involved, rather than a distinction in the nature of these situational dimensions, such as conceptual complexity.

5.2.3.2 The READ component

The results for this component, unlike those of OP, appeared to be relatively unaffected by test method. Word recognition was of considerable importance, in

contrast to the OP text. This may be because lexis with more syllables contained in the OP text was frequent and familiar, but that this was not the case in READ. Lexical access, in the form of word frequency, was important for item difficulty, as it was with OP, but syntactic parsing was not. It may be that the larger stretches of contextualised text meant that comprehension depended less on syntax than with the OP text, since readers could make use of a broader range of information when decoding the text. Other indicators associated with syntax but expected to influence the establishment of propositional meaning more than parsing (X043.READ.fronted.CLPS and X044.READ.passive) were found to be significant, suggesting that non-standard word order is a difficult feature for learners at and around B2. When establishing a coherent textbase, lexical cohesion, rather than explicitly marked cohesion, was found to affect item difficulty. In respect of building a situational model, continuity of causal links within the text were important for item difficulty but intentional and temporal links were not.

Interestingly, the findings of the current study concerning the READ component (see Figure 11 in 4.2) relate closely to those of other studies, such as Wu (2014) and Ilc and Stopar (2014), discussed in 2.2.2. Both studies found that judges believed lower level processes, such as *word recognition* and *lexical access*, were more frequently elicited by items than higher level processes such as *creating a text level structure*. In their research, evidence of the elicitation of all subcomponents was found in FCE items. Both studies concluded that, particularly as FCE's B1 level sister test, PET, exhibited less evidence of higher level processes, these findings were to be expected for a B2 level test.

Due to methodological differences in the current study, a direct comparison of results cannot be made with the findings of Wu (2014) and Ilc and Stopar (2014) (2.2.2). However, a comparison between Figure 11, which summarises the influence of the subcomponents used in the current study according to empirical data, and Figure 2 from the Wu (2014) study shows a similar configuration. Again, a greater influence of lower level processes in the test, although evidence of the entire range of subcomponents (except syntactic parsing) was exhibited. It should

be noted that the absence of syntactic parsing cannot be interpreted as evidence of its lack of importance in the current study, since other measures which were not implemented in the study may have detected it. Its absence does not, therefore, count against the comparison. Instead, the entire graph can be interpreted as circumstantial evidence supporting the conclusions of Wu (2014) and Ilc and Stopar (2014) that the balance between lower and higher level processes is as they found it. In other words, FCE taps lower level processes more than higher level processes but elicits processes from most of the range specified by the Khalifa and Weir (2009) model. This is, perhaps, not surprising, as lower level processes are required before a reader can engage in higher level processes 2.2.

5.2.4 RD

5.2.4.1 RD indicators and subcomponents

For this component, six indicators were tested. Four of them were based on the LSA of options (keys and distractors) and their relevant text (X054.RD.LSA.term.KEY, X055.RD.LSA.term.DIST, X056.RD.LSA.doc.KEY, X057.RD.LSA.doc.DIST), and so, for convenience, were considered as one subcomponent. 'Term' differs from 'doc' (document) in that, for the former, matches between texts are at a word level, whereas for the latter, the semantic whole of the text was considered (see 2.7.3). The two remaining indicators (X058.RD.disperse, X059.RD.pract) were each given separate subcomponents because they shared no particular similarity. Two indicators were found to exert an interpretable influence over item difficulty. Results at subcomponent level are contained in Table 71.

All four effects dealing with the semantic match between option and relevant text were significant and interpretable. In other words, as expected, a good semantic match between key and relevant text makes items easier, and a good semantic match between distractor and relevant texts makes items more difficult. As described in 3.6.1.6, in addition to distinguishing between keys and distractors, the indicators used separated matches based on each term in the text and

matches based on the semantic whole ('doc'). This distinction was hypothesised to replicate the distinction between using the textbase to determine the response decision, and using the situational model (2.7.3). By far the most significant of these four indicators was that for the key document match (X056.RD.LSA.doc.KEY). This suggests two conclusions: first, as confirming the key discriminates between candidates more, the need to confirm the key is of more consequence than the need to reject the distractors. Second, there is a need to use the situational model to confirm the key. In both cases, this is as would be hoped for a reading comprehension test, since the use of distractors is an artefact of the test and the need to use the situational model means that the items are not prone to word spots. The importance of confirming the key contrasts with Embretson and Wetzel's (1987) assertion that confirmation is only important where disconfirmation of options fails. A likely reason for this difference is that the items found in the tests they examined were or poorer quality than those found in FCE, and confirmation of the key was not always necessary. For the indicators concerned with distractors, the term match indicator (X055.RD.LSA.term.DIST) showed most influence, albeit noticeably less than the doc indicator concerning the key. That a term to term match discriminated more for distractors suggests that when disconfirming distractors, the meaning of individual words is most useful, and that a deeper understanding of the sense of the option and its related text was less important.

Scrutiny of Appendix 11: incidence matrix summary reveals that the X056.RD.LSA.doc.KEY indicator is highest for Part 1 and lowest for Parts 3 and 4. Since the aggregate effect of the indicators is of interest in this study, the presence of this variation does not affect overall conclusions but is of interest to provide further information about the underlying phenomena. A likely explanation for the matches in Part 1 being high is that the options consist of headings which are designed to summarise the content of paragraphs. The options for Part 2 largely work by paraphrasing information contained in or implied by the relevant text, but do not summarise all of it, so lower semantic matches are to be expected. The

options for Part 3, by contrast, although maintaining links to the surrounding text, add information to the text, so less semantic overlap is to be expected. Part 4, like Part 2, operates largely on the basis of the options paraphrasing parts of the relevant text. It is therefore somewhat surprising that the semantic overlap between the options and the relevant text is lower than that of Part 3. However, the reason for the difference in coefficient between Parts 2 and 4 is likely to be related to the need for the reader to infer information in Part 2 but simply to locate information in Part 4 (Khalifa & Weir, 2009). After using semantic overlap to locate relevant information in the text, candidates may still need to make an inference to respond correctly in Part 2. In Part 4, there is less need to infer, so item writers make the items sufficiently difficult by reducing semantic overlap.

The indicator measuring the dispersal of information was also found to be interpretable, but with a very small coefficient. Particularly large spacing of information was found in Part 4, whereas Part 3 showed the lowest levels. It may be, therefore, that the influence of this attribute is only felt in some tasks (i.e. Part 4) and not in others (i.e. Part 3), or that such contextual features do not discriminate between candidates at the B2 level.

5.2.4.2 The RD component

The match between the text for the key and its relevant text in the reading passage was found to be the most important factor in discriminating between candidates when they are determining the response decision. Furthermore, as predicted, the match was based on the relationship between the semantic content of each text as a whole, rather than terms within them. This appears to suggest that, i) the confirmation of the key is more important than disconfirmation of distractors, and that ii) at the stage of responding to items, candidates need to use a situational model of the text they have read to confirm the key. These features appear to vary in importance across tasks in relation to task features.

5.3 Research question 4: test method effects

One of the striking findings in this study was the relatively large effect of the OP (net influence: 12.56%) and RD (net influence: 14.66%) components (see Table

72). In both cases, their influence would normally be considered, in large part, construct irrelevant because these effects relate to items, not the process of reading per se (2.7). In other words, as tests often try to facilitate inferences about non-testing situations, where items are not encountered, the influence of items is considered problematic. Nevertheless, some influence is unavoidable, whatever item type, or *test method*. Such questions are of particular interest in test construction or revisions, as the effects may be mitigated to some extent. Also, such significant findings show that test method effect should be investigated in studies of the cognitive processes involved in test taking.

The fourth research question was as follow:

> What evidence can be found of test methods effects influencing item difficulty?

Since each test part is realised with a different test method, this question involves the consideration of each test parts, individually and in comparison with each other. In this section, the influence of fixed effects which indicate the influence of test method effects for specific parts will be discussed. Information about dependency between items is also relevant where this is a test method effect.

5.3.1 Test method effect by test part and for all parts

5.3.1.1 Part 1

In this part, candidates were required to match a heading to a paragraph (Appendix 1: test papers), with headings generally summarise some or all of the paragraph. This is likely to be the cause of a high semantic match between key and relevant text, described in 5.2.4.1. Some item pairs, notably items 1 and 3 were found to have relatively large dependency (4.4.1.1, 4.5.2). This is perhaps not surprising, as the two keys are:

E People are born with certain preferences regarding fitness

H Any methods of keeping fit can be very enjoyable for some people but very unpleasant for others.

Both are quite similar semantically, as are their respective paragraphs. The response method, the fact that each option can be selected once, causes dependency, as using the key for the other item of the pair automatically means that the other item will be answered incorrectly. Items 1 and 3 exhibit the most LD within this task (see 4.4.1.1).

5.3.1.2 Part 2

This part was not highly impacted by high noun modification and left embeddedness according to the Appendix 11: incidence matrix summary. For other test parts, their high impact was explained by the use of incomplete sentences (5.2.3.1.3). Part 2, however, with the exception of item 9, is the only test part with complete sentences for the OP text, if the option is understood as a continuation of the stem for items 11 and 15. In addition, the semantic match between key and relevant text (5.2.4.1) was relatively high. As a result, there was relatively little test method effect of the kinds investigated discovered in this test part.

5.3.1.3 Part 3

As described in 5.2.3.1.3, left embeddedness was relatively more common in this task, in order that a link between the option and the text which surrounds it in the reading passage can be established. In addition, as with Part 1, dependency between items in this part was apparent, particularly 17 and 18 (4.4.1.1, 4.5.2). Also like Part 1, the cause of the dependency was the response format, where the choice of a good distractor for one item automatically meant a second wrong answer if that option was key for another item.

5.3.1.4 Part 4

As argued in 5.2.1.1.3, items in this part are characterised by densely packed information in the item stem. This allows the correct answer to be selected if all the information is matched, but makes near matches possible with distractors. The result is a high level of noun modification, which affects item difficulty. As with Part 1, LD due to the response method was evident. Items 17 and 18 showed the greatest degree of dependency.

5.3.1.5 *All parts*

In addition to the influence of those indicators mentioned above, it was hypothesised in 5.2.1.2 that the importance of OP content words in all parts was due to the reduced grammatical and contextual support available in the item stems and options compared to the reading passage. In other words, the test method of all tasks involved item stems and options which were relatively decontextualised compared to the reading passages, in which a context is built up. This decontextulisation seems to contribute to item difficulty because it places greater emphasis on the lexical and grammatical features of the stem and option text.

5.4 Model: research question 5: variance explained

The final research question concerns the amount of variance explained by the final model. They are:

> What proportion of the variance of the corresponding Rasch model does the LLTM account for?

A model is often judged on its ability to explain the data it is fitted to. Freedle and Kostin (1993), for example, claim 39% of the variance explained by their indicators in a data set comprising data from 213 items. In the current study, the figure was calculated in relation to the variance explained by the Rasch model (3.8.3.1.3). Such a comparison is normal for LLTMs, as they involve the imperfect decomposition of the item difficulty parameter from the Rasch model (2.9.3.3.3). The LLTM explains 75.79% of the variance of the Rasch model with corresponding corrections for violations of dimensionality and local item independence (see 0).

There are several reasons which account for the discrepancy between the amount of variance explained by the LLTM and the Rasch model. The principal aim of the current project was to investigate construct representation of a single form of one test component. This was felt to be important because the results of individual candidates are derived from their performance on from single test forms. An additional aim of the current research is to determine how well such investigations

could be operationalised to investigate single test forms, therefore. This meant that, rather than pool items from many forms for the investigation, the data would be derived from just one. An inevitable consequence is that only the indicators with the strongest influence over the test would be detected. Had a larger sample of items been used, it is likely that more indicators would have been found to be both significant and interpretable. Furthermore, with a larger sample of texts and items, measurement error would have been reduced and the influence of each indicator would have been made clearer. It should also be remembered that the impact of several indicators were shown to vary across test parts (5.3). Such variance effectively subdivides the sample of data used to detect these features and therefore increases measurement error further, and contributing to variance which was unaccounted for.

5.5 Method

One of the aims of the study was to determine the effectiveness and utility of the method used when applied to data from a single test form. The method is clearly effective and useful. In contrast to approaches outlined in the literature (e.g. Weir, 2005; Khalifa and Weir, 2009) and a number of empirical studies (Wu, 2014; Ilc and Stopar) it successfully provides a way to relate contextual features to cognitive processes empirically and to relate this to item difficulty. Unlike research methods, such as eye tracking, simultaneous verbal protocol and stimulated recall, which are used in experimental designs, the current method can be done using live test data, something of great interest to test providers and other researchers involved in validation studies. However, unlike methods employing expert judgement, it cannot claim to explain all variance in the data, which means that further research to uncover further sources of difficulty is always implied. In the case of the current study, although the amount of variance explained by the fixed effects in the final model was relatively large, it is smaller than that explained by the Rasch model, and this unexplained variance leaves room for uncertainty. If this study is replicated, new indicators would need to be tested, in addition to those which

formed part of the current study, to account for the unaccounted for variance. Further research is discussed in more depth in 5.7.

In respect of applying this method to the analysis of a single test form with several test tasks, a significant finding was that several factor based indicators (e.g. those dealing with propositions) did not prove useful. This seems largely due to the small number of items, which made it difficult to determine a usefully descriptive series of levels within the factors. In future small-scale studies, therefore, it may be better to concentrate on indicators based on continuous variables, or to explore what the significant boundaries between levels of each indicator are likely to be.

As suggested by Weir (2013:473), the use of computer generated indices offer 'the potential of a more systematic, efficient way of describing a number of contextual parameters'. The current research bears this out, with 15 computer generated indices found to influence item difficulty in significant and interpretable ways.

5.6 Generalisation and use of findings

It has been shown that the test form used for this research was reasonably representative of the test in general (4.2.2.1). For this reason, similar results would be expected if the same method were applied to any other test form with the same test tasks[14]. The indicators with the largest coefficients would be expected to be found again, although other indicators might not. These features can therefore be understood as the consolidated findings of the current research in respect of FCE. In respect of further research on FCE Reading, or other tests, the indicators found to be significant and interpretable are clear candidates to be included in other research. Those indicators which were not found to be useful may be trialled again but with less expectation of success. In respect of such research, four categories of indicator present themselves to the researcher:

1. indicators with large significant and interpretable values
2. indicators with small significant and interpretable values

[14] FCE Reading was revised in 2008 and Part 1 was dropped. Such a comparison would therefore require a pre-2008 test form.

3. indicators which are either not significant or uninterpretable or both
4. indicators which were not tested in this study

The first category would be expected to be of use in any future study, whereas the second and third are more likely to be dependent on the test form used. The fourth category would be up to future researchers to fill.

The investigation of the indicators for OP and RD were particularly interesting for test providers, as the influence of the option text and the response decision are not typically addressed (e.g. the model of validation proposed by Field (2013)). For example, as discussed in 1.3, reading in a test situation will be influenced by attributes which are not intended to be part of the construct. Investigation of the impact is a necessary first step for test providers to explore ways of mitigating their impact. This is likely to involve balancing their effects, rather than removing them altogether. This is because the nature of standardised testing requires the control of the task which items provide. These effects can be significant, Embretson and Wetzel (1987) concluded that the response decision was more important than the process of text representation. Their model did, however, explain only 37% of the variance in the data, meaning that explanation of the balance could identify other important influences and lead to quite different conclusions.

Findings such as those described in this thesis are likely to be of great interest to test providers in future revisions of FCE. For example, item stems and options are important for the formation of the task model but they are relatively decontextulised. If they are written as incomplete sentences, as in Parts 1, 3 and 4, they appear to become more difficult, and this is likely to be due to two reasons: because less linguistic information is available to interpret them (5.2.1.2), and because they become more densely loaded with information because there is a requirement that the stems are short (5.3). This suggests that test providers should supply more context and relax word limits for item stems or options so that less emphasis is placed on their interpretation. The extent of LD due to test method (the pooling of keys and distractors for all items) in Parts 1 and 3 is also a concern

(5.3). Item types such as multiple-choice are less affected by LD, but may have other disadvantages. An alternative way to deal with this issue when computing results would be to use measurement models which account for LD (Wainer et al., 2007). Just as was described in 3.8.2.1 and 3.8.3.1.1 and shown in 4.4.1 and 4.5.2, LD can be modelled statistically and thereby the effect on the measurement model mitigated.

5.7 Limitations and further study

One clear limitation of the current study is that not all indicators were interpretable, so the status of some attributes is unclear in respect of FCE (0). The reason for this is not easy to determine without replication studies using other data. Likely reasons for the inability to recover information on the rejected indicators are:

- the ability of candidates did not vary significantly on the indicators (Reckase, 2009) – see 2.7
- an insufficient number of items was available to recover adequate information on all indicators, given capitalisation on chance and measurement error – see 2.8

In so far as the first explanation is true, conclusions about specific indicators may be considered applicable to FCE candidates in general. As discussed in 2.7, response data is the product of a particular test form and a particular group of candidates. Evidence presented in this study has shown that the group of candidates who provided responses were broadly representative of typical FCE candidates (4.2.2.2). Furthermore, influence on the data due to a single L1 was mitigated during the sampling process (3.3.2). The test materials were also shown to be relatively typical (4.2.3), however, only thirty-five items were included in this study. The complexity and variety of language means that, although the items reasonably sample the intended linguistic domain, it is possible that a specific single feature (such as a word) would be very influential in one test form but not appear in another. The results of this study, therefore, are at some risk of

capitalisation on chance. It may be the case that, due to the specific language used in another test form, some indicators would obtain quite different results in a replication study.

One way to lessen capitalisation on chance would be to conduct a similar study with data from more than one test form, although such an approach would not provide answers to all questions. Such a study would mean that the pooled data would represent a generalised construct of FCE and the influence of particular linguistic items would be reduced. An additional benefit of a study involving more data would be in its identification of key indicators which would be more readily generalisable across test forms. Such studies would, however, also need to account for attributes which were specific to only some test forms, as attributes which would impact on the interpretation of results for some forms may not be detected. A study, like the current one, involving a single form could be used to do this. For this reason, multi-form studies must be seen as complimentary to single-form studies. A single form was used for the current study because operationalising a method of construct validation for single test form was of specific interest here (1.7).

Measurement error is also mentioned above as a possible reason lack of interpretable information about some indicators. Error cannot be discounted from the process of obtaining the initial indices from which the indicators were constructed. This may be systematic error in the machine-based indices, or random and systematic error in the human judgement process (2.9.2). Error can never be eradicated, but ways to mitigate error may be found. The software providers who are responsible for facilitating the machine-based indices do not, in general supply very extensive information about error in their systems, although confidence intervals are available for the software identifying propositions (Brown et al., 2012; Covington, 2012). There is also relatively limited information about the precise way in which many of the indices are calculated (Weir, 2013). This means that, for the more complex indices at least, the way which they are operationalised may contain assumptions or compromises which makes them less

accurate for the researcher's purposes. This will, however, remain unknown to the researcher.

As with indices constructed using computer measurement, human judgement is not infallible and there are many ways in which error could be introduced into a judgement process. For example, lack of clarity in the instructions, fatigue and time factors. Ways to mitigate such effects are available in an extensive literature on rating in educational measurement. These include practice sessions, collecting judgements from a larger group, or organising more rounds of judgement and discussion. Such additions to the process, however, come at a cost of the requirement for more time and effort from the judges and these were not feasible in the current study.

The current study could be usefully extended in a number of ways. In addition to studies involving more test forms, as mentioned above, the effect of attributes at different ability levels would be of interest, for example, to provide a perspective on latent growth, and to determine criterial features at different levels. Such research could lead to a revised version of the CEFR (Council of Europe, 2001), which is in turn an extremely useful tool in many areas of language education, including in assessment. This could be done by replicating the current study with a range of other tests at different levels. FCE is one of five tests which form a suite covering levels A2 to C2. Further interesting research could be carried out to determine patterns with which attributes influence difficulty across this scale. For example, some attributes may be found important at some levels but not others. It would make sense to first investigate tests in the same suite because significant differences between these tests are more likely to be the result of ability levels, than other causes.

A study focussing on construct representation of different tests at the same level would yield information concerning how the constructs of the two tests differ. Such tests might be two tests of general language ability, like FCE, or could differ in test purpose. It would for example, be of interest to determine whether a test

a Language for Academic Purposes (LAP), carries fundamentally different influences on test difficulty than a general purpose test.

Also of interest would be a mixed-methods approach incorporating a similar quantitative component to that of the current study with a qualitative component involving a smaller-scale study with eye-tracking or verbal protocols as a methodology. The latter could help to identify attributes which could then be the subject of a large scale study, using an analytical methodology like the current one, to determine their actual influence on item difficulty.

5.8 Implications of the research for specific groups

The current research is likely to be of interest to a number of groups, including the developers of FCE, other test providers, researchers wishing to employ the Khalifa and Weir (2009) model of reading and researchers intending to use the procedures developed for the current study. The implications of the research, as it pertains to each group, will therefore be set out in this section.

5.8.1 Developers of FCE

The findings of this study have clear implications for the developers of FCE. They must consider the extent to which the construct representation revealed through this study matches what was intended. Where it does not, changes to the test are implied. It must be noted however, that the current study examines data from a single form of the FCE Reading component, dating from 2005. Some indicators may be relevant to the 2005 test but not to others (see 5.7). Before making changes to the way the test is structured or designed, therefore, it would be important to replicate the study on a range of other data.

Notwithstanding the need to compare any differences between the intended construct and that revealed through this study, the findings relating to test method effect are almost certainly unintended (see 5.3). The findings in question relate first to the difficulties in syntactic parsing of the option and stem text. They were not a feature of the reading passage, so do not appear to be simply a feature of reading that text. Left-embeddedness has a particularly strong effect for Part 1

and Part 3 items, and noun modification for Part 4 items. The presence of LD was also found in Parts 1 and 3.

The influence of test method effects is, however, not straightforward to remove. As is clear from Rouet's (2012) model, as well as the composite model employed in the current study (see 2.3 and 2.5), item text is used to form the task model in order to set reading goals and to monitor them. For this reason, artefacts of the test situation relating to item text are always likely to be present, where they would be absent in non-test situations. One option open to test developers, therefore, is to employ a range of item types (in other words continue with the current FCE format). Having a mix of task types at least, means that test method effects in one task may be compensated for by their absence from another.

Implications of the findings concerning test method effects for tests that include multiple task types include consideration of whether there is any way to mitigate the effects found to influence difficulty, and whether the balance between the effects manifest across tasks is appropriate. It should be noted here again that Part 1 was dropped from the test in 2008. For this reason, left-embedeness has a major effect on only one part of the test, so the test developers may decide that balance of effects is appropriate. An example of the way in which test developers might mitigate effects concerns Part 4. If unintended consequences could be ruled out (for example, increased reading time), the length of the item stems could be increased to avoid noun modification affecting difficulty in that task.

The results concerning RD were as expected for a test of reading comprehension. This study found that being able to form a situational model of key and the relevant text and selecting the correct response based on their match discriminated between candidates (5.2.4). Understanding distractors was relevant but made much less of a difference to performance. Embretson and Wetzel's (1987) study appears to find that a strategy of disconfirming options was viable for the tests they examined however. The findings associated with RD for FCE,

however, unlike those for the test in the Embretson and Wetzel (1987) test, do not imply a need for change.

5.8.2 Other test providers

Other test providers should investigate their tests in a similar fashion to that laid out here. This will help them to determine the extent to which the intended construct is represented in their tests, and thereby lead them to consider which, if any, actions need to be taken. Those test providers responsible for developing B2 level tests of English should also investigate whether criterial features at that level are represented in the construct. If a test is to be considered B2, it would be expected to distinguish between candidates on such features, and this would be a way in which to compare tests from different providers.

In the current study, frontedness and the passive voice were found to influence the difficulty of the reading passage, but syntactic parsing indicators with standard word order were not (see 5.2.3.1.3 and 5.2.3.1.4). This appear to corroborate Hawkins and Buttery's (2012) research, suggesting that pseudo-cleft sentences were among the features which differentiated B1 from B2. In other words, in the reading passage, candidates are challenged by non-standard word order at this level, but not by syntax with more familiar word order patterns. For this reason, indicators of non-standard word order could be considered criterial and should be investigated in any tests at this level.

In addition to ensuring that the intended construct is represented by the test results, just as with the developers of FCE, it is important for test providers to ensure that test method effects are minimised and balanced across the test. Furthermore, the RD process should favour candidates who select responses based on their reading of the passage, rather than those whose strategy prioritises matching words in options with those in the passage, with the hope of disconfirming the options.

5.8.3 Researchers wishing to employ the Khalifa and Weir (2009) model of reading

Researchers employing the Khalifa and Weir (2009) model are advised not to do so without modifications. Specifically, as discussed in 2.2.3, the inferencing stage should be reconsidered, as inferencing is not essential to pass from establishing propositional meaning to building a mental model. According to the work of Kintsch and van Dijk (1978) among others, there is, however, a need to establish a coherent textbase, and this may require the use of bridging inferences in particular. For this reason, the stage has been treated as establishing a coherent text base, rather than simply inferencing. The difference is not merely the label, as the selection of indicators for each stage, and the interpretation of their coefficients depends on how each stage is defined.

Another area of weakness in the Khalifa and Weir (2009) model of reading is that of goal setter and monitor. As discussed in 2.2.3, the way in which contextual features influence the cognitive process through the goal setter is not fully explained. This has led to a scheme for validation of tests, laid out in Weir (2005) and implemented by researchers such as Wu (2014) and Ilc and Stopar (2014), where cognitive processes and contextual features are investigated separately and the effect of the latter on the former is not part of the process of empirical validation (see, for example, 2.2.2). The current study shows how the two strands may be brought together, using response data.

Related to the discussion of the approach typically implemented in validation studies using the Khalifa and Weir (2009) model of reading, is the potential for further improvement through adoption of Embretson's (1983) conceptualisation of construct representation (see 1.2.2), where the influence of the construct on test results is considered of primary importance. Wu (2014) and Ilc and Stopar (2014), in addition to investigating cognitive processes and contextual feature separately, did not verify which processes and features actually discriminated between candidates taking the test, and thereby impacted on the test results.

Some elements they identified may, therefore, be redundant when considering what is being tested.

Finally, in the approach to test validation associated with the Khalifa and Weir (2009) model, and articulated most clearly by Field (2013), it is difficult to determine the influence of contextual features associated with the task setting, such as test method, or order of items. This is probably because the Khalifa and Weir (2009) model is designed to describe the reading of expert native speakers in non-test situations. A model of reading in test situations, such as the composite model employed in the current study (2.5), is useful in such situations as it explains the way in which features which do not appear in non-test situations, such as the text of items and the way in which the response is determined, influence the reading process. Given the influence these features were found to have (4.2), they should be investigated as a matter of course.

5.8.4 Researchers intending to employ the procedures developed for the current study

The implications of the current research for those wishing to adopt a similar approach relate mainly to the way in which they choose to frame their study. Clearly they may have somewhat different purposes form those of the current study, and they will have different data. The selection of indicators is a key element of any such study. Future studies could take the indicators found to be interpretable here as a starting point. Section 5.6 suggests that indicators may be divided into four categories on the basis of the current study. Those in the first category (indicators with large significant and interpretable values) are clearly prime candidates for use in future studies. However, as the discussion on criterial features at B2 implies (5.8.2), not all indicators will be productive at other levels. Indicators relating to non-standard word order, for example, may be found not to discriminate at relatively lower (A1/2) or higher (C1/C2) levels even though they have proved criterial at B1/B2.

The methodology of the current study might also be used, suitably adapted, for a variety of different aims. In the current study, the construct representation of a

single test form was of principal importance. It could be that, in other studies, something more general than a single test form would be of interest (e.g. the test in general, as characterised by multiple forms). On the other hand, the focus might be more specific than the representation of the entire construct (e.g. test method effects found in a single task). In these cases, the data would clearly be gathered from as many forms as was available but the indicators would be limited just to those concerning test method effects.

5.9 Achievements of the current study

The current study has developed an innovative approach to investigating test constructs for language testing. The novelty is fivefold:

- unlike some other investigations into cognitive processing in language testing (Ilc & Stopar, 2014; Wu, 2014), it features a link between item difficulty and cognitive processes (construct representation)
- also unlike some other investigations into cognitive processing in language testing (Ilc & Stopar, 2014; Wu, 2014) and those suggesting approaches to validation (Field, 2013), a link between contextual features and cognitive processing is made
- the importance of the influence of 'construct-irrelevant' contextual features (such as test method effects) is shown. This implies that they should be investigated alongside other contextual features
- a sophisticated statistical modelling approach was adapted
- machine generated indices were employed to extract information about attributes of the test materials

The first two points are particularly important. The primal literature on cognitive processing in language testing (Khalifa & Weir, 2009; Weir, 2005) does not describe how, or even suggest, that those parts of the process which make tests harder can be distinguished from those processes which are necessary to

complete a task but do not discriminate between candidates. Furthermore, although it is made clear that there is a link between contextual features and cognitive processes, the way to relate the two directly is not elaborated. Linking cognitive processes with item difficulty and contextual features with cognitive processes is a fundamental part of this study.

As with the first two points, the investigation of 'construct-irrelevant' contextual effects has been overlooked in previous language testing research. The current study shows the utility of investigating these effects alongside 'construct-relevant' contextual features.

In respect of the fourth point above, current approaches to examining cognitive processing in language testing tend to make more use of qualitative than of quantitative procedures. The result is usually that caveats about the motivation, preparedness and representativeness of candidates in the study must be made, as data cannot come from live test performance. The current study is different, as all results are based on the performance of a considerable number of live test candidates, therefore these caveats are not required.

The final point, concerning the use of machine generated indices, is not unique (Aryadoust & Goh, 2014; Weir, 2013; Wu, 2014). Their use is, however, becoming more common due to its utility. It is important, therefore, to determine which indices may be useful and why. The current studies trials more than 50. This is significant not only for the construct representation of FCE Dec 2005, but also for the construction of future test forms and the revision of the test. Furthermore, knowing more about the importance of specific test attributes offers a foundation for research into the automatic item generation.

The findings of the current study also contain several important lessons for research into test constructs. First, it argues that the investigation of construct representation (the relationship of contextual features and cognitive processes to item difficulty) is important and shows that it is feasible. Second, the importance of test method effects in construct representation have been demonstrated.

Studies in this area should, therefore, aim to investigate test method effects as part of construct representation. Finally, the study demonstrates the potential of interrogating live test data for investigations of cognitive processes. This is important, as many studies focus only on qualitative methods which require an experimental design.

The approach in the current study has facilitated the investigation of FCE Reading. The study presents specific information on the construct of this test and makes suggestions which may be taken into account in future revisions. In short, the current study has achieved all the goals set out for it in 1.7:

- To determine elements of the construct representation of the Reading paper of a form of First Certificate in English (FCE) administered in December 2005 (FCE Dec 2005).
- To develop a practical method which can be deployed in the construct investigation of reading tests with varying test methods.
- To trial the use of machine generated indices in the construct investigation of reading tests.

Appendix 1: test papers

UNIVERSITY OF CAMBRIDGE ESOL EXAMINATIONS

English for Speakers of Other Languages

FIRST CERTIFICATE IN ENGLISH **0100/1**

PAPER 1 Reading

[Day] **December 2005** Morning 1 hour 15 minutes

Additional materials:
Answer sheet
Soft clean eraser
Soft pencil (type B or HB is recommended)

TIME 1 hour 15 minutes

INSTRUCTIONS TO CANDIDATES

Do not open this booklet until you are told to do so.

Write your name, Centre number and candidate number on the answer sheet in the spaces provided unless this has already been done for you.

There are thirty-five questions on this paper.

Answer **all** questions.

For each question (**1-35**), mark one answer only.

Mark your answers on the separate answer sheet. Use a soft pencil.

INFORMATION FOR CANDIDATES

Questions **1-22** carry two marks.

Questions **23-35** carry one mark.

Part 1

You are going to read an article about fitness and exercise. Choose from the list **A-I** the sentence which best summarises each part (**1-7**) of the article. There is one extra sentence which you do not need to use. There is an example at the beginning (**0**).

Mark your answers **on the separate answer sheet.**

A	Some people stop trying to keep fit because of a false impression they have of others.
B	People have been discouraged from doing the kind of exercise they would find enjoyable.
C	It is wrong to see exercise as an obligation.
D	People wrongly believe that there are a limited number of ways of keeping fit.
E	People are born with certain preferences regarding fitness.
F	Some people never find a method of keeping fit that suits them.
G	It is wrong to think of exercise mainly in terms of the physical changes it causes.
H	Any methods of keeping fit can be very enjoyable for some people but very unpleasant for others.
I	A theory has been developed to explain why so many people stop bothering to take exercise.

[Turn over

Want to get fit? Do what you like!

0	I

Most good intentions to get fit are forgotten within the first year, while 60 per cent of gym members give up less than six months after paying their registration fees. If you have ever wondered why you can't stick at keeping fit, the answer, according to the leading American fitness expert Peg Jordan, could be in your genes.

1	

In her latest book, *The Fitness Instinct*, Jordan suggests that each person has a natural attraction towards certain types of exercise, and that, while some of us are genetically programmed to run, others prefer less demanding routes to staying in shape.

2	

As part of her mission to find out why there are so many exercise drop-outs, the American expert interviewed almost 2,000 people over four years. What she discovered was that most of the people simply didn't know what activity best suited their personality. 'For years,' says Jordan, 'the fitness industry has persuaded us to ignore our personal preferences. Exercise has been made too scientific, too dull. We should get back to having fun.'

3	

Not only does the type of person we are affect what we like to do, but also each form of exercise will produce widely varying mood responses in different people. According to Jordan, when people choose the right exercise, their body responds with a feel-good sensation or 'exercise high'. Forcing yourself along to aerobics when it doesn't suit your natural tendencies, on the other hand, will feel little short of punishment. That is why the same fitness class can leave one participant on an exercise high and another swearing that it is the worst thing they have ever done.

4	

Some British experts agree with Jordan's theory. 'Very often the problem is that people join a gym because they assume that is where they should get fit rather than take time to think about what kind of exercise they might really enjoy,' says Dearbhla McCullough, a sports psychologist. 'In a way, we have become socially programmed to try only certain types of exercise. What we like doing, however, is based on the kind of people we are – our personalities stimulate our hobbies and habits. Choose a fitness routine you don't really enjoy, and before long you will lose interest.'

5	

Another problem is something that Jordan terms the 'intimidation factor'. She noticed how the majority of women claimed they were discouraged from working out regularly because their bodies looked nothing like the perfect figures of the stars who appear in the glossy magazines. 'They give up before they've started. They pick up a magazine, look at a photograph and feel they'll never have that body, so why try?' she says. 'They decide it's not worth the effort. What they don't realise is that these people spend 30 hours or more a week with their personal trainers and, when all else fails, have artists alter the photograph.'

6	

Jordan suggests the answer to stop-start fitness plans is to take a long, hard look within. 'Ask yourself some questions about what it is you like and hate about the gym, for instance. And then stop thinking about exercise in terms of calories burnt and changes in body-fat percentage, but as fun,' she says. 'The path to fitness is a self-awareness path. The more you know about yourself and the more you accept your body as it is, the more movement becomes a daily act.'

7	

There are plenty of times, says Jordan, when you will finish a hard day at work and ask yourself: 'Should I really go to the gym?' Or, after a late night out, you may wake up tired but think: 'I should go for a run.' 'Thinking of exercise as a "should" leaves you with only two choices,' she says. 'If you don't exercise, you not only blame yourself for making the wrong choice, you also strengthen the negative pattern of "should but didn't" and label yourself a failure.'

[Turn over

4

Part 2

You are going to read a newspaper article about a musical family. For questions **8-15**, choose the answer (**A, B, C** or **D**) which you think fits best according to the text.

Mark your answers **on the separate answer sheet.**

Meet the Amazing Watkins Family

The sons are composers and prize-winning musicians, while Dad makes the instruments. ***Matthew Rye*** *reports.*

Whole families of musicians are not exactly rare. However, it is unusual to come across one that includes not only writers and performers of music, but also an instrument maker.

When South Wales schoolteachers John and Hetty Watkins needed to get their ten-year-old son, Paul, a cello to suit his blossoming talents, they baulked at the costs involved. 'We had a look at various dealers and it was obvious it was going to be very expensive,' John says. 'So I wondered if I could actually make one. I discovered that the Welsh School of Instrument Making was not far from where I lived, and I went along for evening classes once a week for about three years.'

'After probably three or four goes with violins and violas, he had a crack at his first cello,' Paul, *line 17* now 28, adds. 'It turned out really well. He made me another one a bit later, when he'd got the hang of it. And that's the one I used right up until a few months ago.' John has since retired as a teacher to work as a full-time craftsman, and makes up to a dozen violins a year – selling one to the esteemed American player Jaime Laredo was 'the icing on the cake'.

Both Paul and his younger brother, Huw, were encouraged to play music from an early age. The piano came first: 'As soon as I was big enough to climb up and bang the keys, that's what I did,' Paul remembers. But it wasn't long before the cello beckoned. 'My folks were really quite keen for me to take up the violin, because Dad, who played the viola, used to play chamber music with his mates and they needed another violin to make up a string trio. I learned it for about six weeks but didn't take to it. But I really took to the character who played the cello in Dad's group. I thought he was a very cool guy when I was six or seven. So he said he'd give me some lessons, and that really started it all off. Later, they suggested that my brother play the violin too, but he would have none of it.'

'My parents were both supportive and relaxed,' Huw says. 'I don't think I would have responded very well to being pushed. And, rather than feeling threatened by Paul's success, I found that I had something to aspire to.' Now 22, he is beginning to make his own mark as a pianist and composer.

Meanwhile, John Watkins' cello has done his elder son proud. With it, Paul won the string final of the *BBC Young Musician of the Year* competition. Then, at the remarkably youthful age of 20, he was appointed principal cellist of the BBC Symphony Orchestra, a position he held, still playing his father's instrument, until last year, when he left to concentrate more on chamber music. Now, however, he has acquired a Francesco Rugeri cello, on loan from the Royal Academy of Music. 'Dad's not said anything about me moving on, though recently he had the chance to run a bow across the strings of each in turn and had to admit that my new one is quite nice! I think the only thing Dad's doesn't have – and may acquire after about 50-100 years – is the power to project right to the back of large concert halls. It will get richer with age, like my Rugeri, which is already 304 years old.'

Soon he will be seen on television playing the Rugeri as the soloist in Elgar's Cello Concerto, which forms the heart of the second programme in the new series, *Masterworks*. 'The well-known performance history doesn't affect the way I play the work,' he says. 'I'm always going to do it my way.' But Paul won't be able to watch himself on television – the same night he is playing at the Cheltenham Festival. Nor will Huw, whose String Quartet is receiving its London premiere at the Wigmore Hall the same evening. John and Hetty will have to be diplomatic – and energetic – if they are to keep track of all their sons' musical activities over the coming weeks.

[Turn over

8 Why did John Watkins decide to make a cello?

A He wanted to encourage his son Paul to take up the instrument.
B He was keen to do a course at the nearby school.
C He felt that dealers were giving him false information.
D He wanted to avoid having to pay for one.

9 What is meant by 'crack' in line 17?

A attempt
B plan
C shock
D period

10 What do we learn in the third paragraph about the instruments John has made?

A He considers the one used by Jaime Laredo to be the best.
B He is particularly pleased about what happened to one of them.
C His violins have turned out to be better than his cellos.
D It took him longer to learn how to make cellos than violins.

11 Paul first became interested in playing the cello because

A he admired someone his father played music with.
B he wanted to play in his father's group.
C he was not very good at playing the piano.
D he did not want to do what his parents wanted.

12 What do we learn about Huw's musical development?

A His parents' attitude has played little part in it.
B It was slow because he lacked determination.
C His brother's achievements gave him an aim.
D He wanted it to be different from his brother's.

13 What does Paul say about the Rugeri cello?

A His father's reaction to it worried him.
B The cello his father made may become as good as it.
C It has qualities that he had not expected.
D He was not keen to tell his father that he was using it.

14 What does Paul say about his performance of Elgar's Cello Concerto?

A It is less traditional than other performances he has given.
B Some viewers are likely to have a low opinion of it.
C He considers it to be one of his best performances.
D It is typical of his approach to everything he plays.

15 What will require some effort from John and Hetty Watkins?

A preventing their sons from taking on too much work
B being aware of everything their sons are involved in
C reminding their sons what they have arranged to do
D advising their sons on what they should do next

[Turn over

Part 3

You are going to read an article about a bird called the kingfisher. Eight sentences have been removed from the article. Choose from the sentences **A-I** the one which fits each gap (**16-22**). There is one extra sentence which you do not need to use. There is an example at the beginning (**0**).

Mark your answers **on the separate answer sheet.**

The kingfisher

Wildlife photographer Charlie James is an expert on the kingfisher: a beautiful blue-green bird that lives near streams and rivers, feeding on fish.

Old trees overhang the stream, half shading shallow water. Soft greens, mud browns and the many different yellows of sunlight are the main colours, as soft as the sounds of water in the breeze. **0** **I** It has gone in a split second, but a trace of the image lingers, its power out of proportion to its size.

Charlie James fell in love with kingfishers at an early age. **16** ____ After all, it is the stuff of legend. Greek myth makes the kingfisher a moon goddess who turned into a bird. Another tale tells how the kingfisher flew so high that its upper body took on the blue of the sky, while its underparts were scorched by the sun.

17 ____ For despite the many different blues that appear in their coats, kingfishers have no blue pigment at all in their feathers. Rather, the structure of their upper feathers scatters light and strongly reflects blue.

18 ____ It's small wonder that some wildlife photographers get so enthusiastic about them. Couple the colours with the fact that kingfishers, though shy of direct human approach, can be easy to watch from a hideout, and you have a recipe for a lifelong passion.

Charlie James's first hideout was an old blanket which he put over his head while he waited near a kingfisher's favourite spot. **19** ____ But it took another four years, he reckons, before he got his first decent picture. In the meantime, the European kingfisher had begun to dominate his life. He spent all the time he could by a kingfisher-rich woodland stream.

The trouble was, school cut the time available to be with the birds. So he missed lessons, becoming what he describes as an 'academic failure'. **20** ____

At 16, he was hired as an advisor for a nature magazine. Work as an assistant to the editor followed, then a gradual move to life as a freelance wildlife film cameraman. What he'd really like to do now is make the ultimate kingfisher film. **21** ____ 'I'm attracted to the simple approach. I like to photograph parts of kingfisher wings …'

The sentence trails off to nothing. He's thinking of those colours of the bird he's spent more than half his life getting close to, yet which still excites interest. **22** ____ But, as Charlie knows, there's so much more to his relationship with the kingfisher than his work can ever show.

[Turn over

A This is why a kingfisher may appear to change from bright blue to rich emerald green with only a slight change in the angle at which light falls on it.

B But his interest in this, the world's most widespread kingfisher and the only member of its cosmopolitan family to breed in Europe, was getting noticed.

C A sure sign of his depth of feeling for this little bird is his inability to identify just what it is that draws him to it.

D The movement sends a highly visible signal to rivals, both males and females, as it defends its stretch of water against neighbours.

E The bird came back within minutes and sat only a metre away.

F The photographs succeed in communicating something of his feelings.

G 'No speech, just beautiful images which say it all,' he says.

H There is some scientific truth in that story.

I The bird cuts like a laser through the scene, straight and fast, a slice of light and motion so striking you almost feel it.

[Turn over

Part 4

You are going to read a magazine article in which various people talk about their jobs. For questions **23-35**, choose from the people (**A-D**). The people may be chosen more than once. There is an example at the beginning (**0**).

Mark your answers **on the separate answer sheet**.

Which person says their job involves

selecting materials to put on show?	**0**	**A**
training high-level staff in their area of work?	**23**	
taking measures to protect public safety?	**24**	
accepting certain financial limitations?	**25**	
encouraging visitor participation?	**26**	
listening to disagreements?	**27**	
doing considerable background research?	**28**	
introducing problems that require solutions?	**29**	
balancing supply and demand?	**30**	
producing advertising literature?	**31**	
organising trips designed to increase people's awareness?	**32**	
constant updating of their own materials?	**33**	
corresponding with the public?	**34**	
working in an area that has personal meaning for them?	**35**	

[Turn over

9

My line of work

Four people talk about their jobs.

A Lisa – Exhibition Programmes Organiser, Science Museum

I'm responsible for putting temporary exhibitions together. This includes planning and designing the exhibition and promoting it. I have to read up about the subject of the exhibition beforehand and then talk to important people in the area so that I can establish the main themes and aims of the exhibition, and plan what objects and pictures should be displayed. I have to make sure the public can understand the thinking behind the exhibition, which means planning interactive displays, workshops and theatre. I also have to bring in engineers and electricians to make sure the final display is not dangerous to visitors. Before the exhibition opens, I help design and write the brochures and leaflets that we'll use to tell people about it.

B Janet – Teacher of London Taxi Drivers

The first thing I do when I get here at 7.30 a.m. is check the accounts. Then I see what new maps and documents need to be produced in order to learn the 'runs' or routes necessary to pass the London taxi-driver test. By midday, about 50 students are in school, working out how to make the journeys. They work out the most direct route, using the correct one-way streets, and right- and left-hand turns. I get involved when there's a difference of opinion – like whether you can do a right turn at a particular junction. When they're close to the test, I'll give them a simple route and no matter what way they say they'll go, I'll tell them they have to use another route because the road is closed. The next student will have to find a third route and again I'll come up with a reason why they can't go that way. It's just to make them think.

C Sarah – Marine Conservationist

I live by the coast and work from home. This involves responding to telephone enquiries, producing educational resources and setting up training courses. Occasionally, I go into our main office but generally I am on the coast. I also work with schools and study centres and run courses for coastal managers and those involved in making decisions about the fate of the seas. I do things like take them out to sea in a boat in an attempt to make them think more about the life underneath them. This often changes their views as it's very different from making decisions using a computer screen. I am extremely lucky because conservation is my hobby, so the job has many highs for me. The downside of the job is that I work for a charity, so there is a constant need for more money. This means I'm always looking for more resources and I'm not able to achieve everything I want.

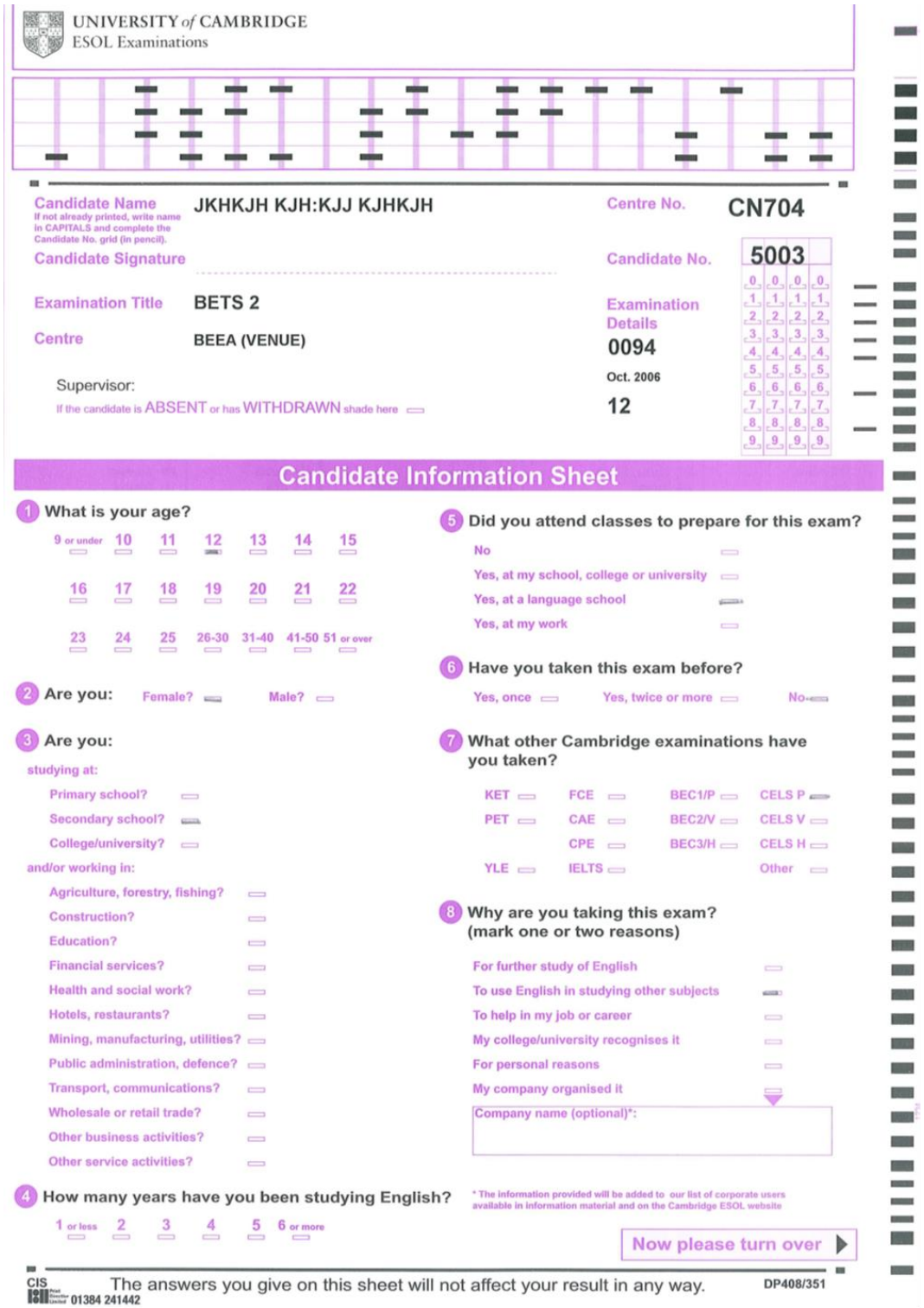
D Chris – Map and Atlas Publisher

My work is pretty varied. I have to make sure that the publishing programme matches market requirements, and ensure that we keep stocks of 300 or so of the books that we publish. We have very high standards of information and content. We receive many letters from readers on issues such as the representation of international boundaries and these in particular require a careful response. I discuss future projects and current sales with co-publishers. I work as part of an enthusiastic group which makes the job that much more enjoyable. The negative side, as with many jobs, is that there is far too much administration to deal with, which leaves less time to work on the more interesting tasks such as product development and design.

Appendix 2: key

Item number, test	Item number, task	Key
1	1	E
2	2	B
3	3	H
4	4	D
5	5	A
6	6	G
7	7	C
8	1	D
9	2	A
10	3	B
11	4	A
12	5	C
13	6	B
14	7	D
15	8	B
16	1	C
17	2	H
18	3	A
19	4	E
20	5	B
21	6	G
22	7	F
23	1	C
24	2	A
25	3	C
26	4	A
27	5	B
28	6	A
29	7	B
30	8	D
31	9	A
32	10	C
33	11	B
34	12	D
35	13	C

Appendix 3: candidate background information form



UNIVERSITY of CAMBRIDGE
ESOL Examinations

Candidate Name
If not already printed, write name in CAPITALS and complete the Candidate No. grid (in pencil).
JKHKJH KJH:KJJ KJHKJH

Centre No. CN704

Candidate Signature

Candidate No. 5003

Examination Title BETS 2

Examination Details 0094

Centre BEEA (VENUE)

Oct. 2006

Supervisor:

If the candidate is ABSENT or has WITHDRAWN shade here

12

Candidate Information Sheet

1 What is your age?

9 or under | 10 | 11 | 12 | 13 | 14 | 15

16 | 17 | 18 | 19 | 20 | 21 | 22

23 | 24 | 25 | 26-30 | 31-40 | 41-50 | 51 or over

2 Are you: Female? Male?

3 Are you:

studying at:

Primary school?
Secondary school?
College/university?

and/or working in:

Agriculture, forestry, fishing?
Construction?
Education?
Financial services?
Health and social work?
Hotels, restaurants?
Mining, manufacturing, utilities?
Public administration, defence?
Transport, communications?
Wholesale or retail trade?
Other business activities?
Other service activities?

4 How many years have you been studying English?

1 or less | 2 | 3 | 4 | 5 | 6 or more

5 Did you attend classes to prepare for this exam?

No
Yes, at my school, college or university
Yes, at a language school
Yes, at my work

6 Have you taken this exam before?

Yes, once | Yes, twice or more | No

7 What other Cambridge examinations have you taken?

KET | FCE | BEC1/P | CELS P
PET | CAE | BEC2/V | CELS V
CPE | BEC3/H | CELS H
YLE | IELTS | Other

8 Why are you taking this exam? (mark one or two reasons)

For further study of English
To use English in studying other subjects
To help in my job or career
My college/university recognises it
For personal reasons
My company organised it

Company name (optional)*:

* The information provided will be added to our list of corporate users available in information material and on the Cambridge ESOL website

Now please turn over

CIS 01384 241442

The answers you give on this sheet will not affect your result in any way.

DP408/351

9 Where do you come from?

If it is not listed, please complete the box 'other'.

001 **Afghanistan**
002 Albania
003 Algeria
004 American Samoa
005 Andorra
006 Angola
007 Antigua
008 Argentina
009 Armenia
010 Australia
011 Austria
213 Azerbaijan
012 **Bahamas**
013 Bahrain
014 Bangladesh
015 Barbados
016 Belarus
017 Belgium
018 Belize
019 Benin
020 Bermuda
021 Bhutan
022 Bolivia
023 Bosnia-Herzegovina
024 Botswana
025 Brazil
026 British Virgin Islands
027 Brunei
028 Bulgaria
029 Burkina Faso
030 Burundi
031 **Cameroon**
032 Canada
033 Cape Verde
034 Cayman Islands
035 Central African Republic
036 Chad
037 Chile
038 China (People's Republic)
039 Colombia
040 Comoros
041 Congo
042 Costa Rica
043 Croatia
044 Cuba
045 Cyprus
046 Czech Republic
047 **Denmark**
048 Djibouti
049 Dominica
050 Dominican Republic
051 **Ecuador**
052 Egypt
053 El Salvador
054 Equatorial Guinea
055 Estonia
056 Ethiopia
057 **Faeroe Islands**
058 Fiji
059 Finland
060 France
061 French Guiana
062 French Polynesia
063 **Gabon**
064 Gambia
065 Georgia
066 Germany
067 Ghana
068 Gibraltar
069 Greece
070 Greenland
071 Grenada
072 Guadaloupe
073 Guam
074 Guatemala
075 Guinea
076 Guinea-Bissau
077 Guyana
078 **Haiti**
079 Honduras
080 Hong Kong
081 Hungary
082 **Iceland**
083 India
084 Indonesia
085 Iran
086 Iraq
087 Ireland
088 Israel
089 Italy
090 Ivory Coast
091 **Jamaica**
092 Japan
093 Jordan
094 **Kampuchea** (Cambodia)
095 Kazakhstan
096 Kenya
097 Korea, North
098 Korea, South
099 Kuwait
100 **Laos**
101 Latvia
102 Lebanon
103 Lesotho
104 Liberia
105 Libya
106 Liechtenstein
107 Lithuania
108 Luxembourg
109 **Macao**
110 Madagascar
111 Malawi
112 Malaysia
113 Maldives
114 Mali
115 Malta
116 Marshall Islands
117 Martinique
118 Mauritania
119 Mauritius
120 Mexico
121 Moldova
122 Monaco
123 Mongolia
124 Montserrat
125 Morocco
126 Mozambique
127 Myanmar
128 **Namibia**
129 Nauru
130 Nepal
131 Netherlands
132 Netherlands Antilles
133 New Caledonia
134 New Zealand
135 Nicaragua
136 Niger
137 Nigeria
138 Niue (Cook Island)
139 Norway
140 **Oman**
141 **Pakistan**
142 Palestine
143 Panama
144 Papua New Guinea
145 Paraguay
146 Peru
147 Philippines
148 Poland
149 Portugal
150 Puerto Rico
151 **Qatar**
152 **Reunion**
153 Romania
154 Russia
155 Rwanda
156 **San Marino**
157 Sao Tome and Principe
158 Saudi Arabia
159 Senegal
160 Seychelles
161 Sierra Leone
162 Singapore
163 Slovakia
164 Slovenia
165 Solomon Islands
166 Somalia
167 South Africa
168 Spain
169 Sri Lanka
170 St.Helena
171 St. Kitts-Nevis-Anguilla
172 St.Lucia
173 St.Pierre and Miquelon
174 St.Vincent and the Grenadines
175 Sudan
176 Surinam
177 Swaziland
178 Sweden
179 Switzerland
180 Syria
181 **Tahiti**
182 Taiwan
183 Tanzania
184 Thailand
185 Togo
186 Tokelau
187 Tonga
188 Trinidad and Tobago
189 Tunisia
190 Turkey
191 Turks and Caicos Islands
192 Tuvalu
193 **Uganda**
194 United Arab Emirates
195 Ukraine
196 United Kingdom
197 Uruguay
198 US Virgin Islands
199 USA
200 Uzbekhistan
201 **Vanuatu**
202 Vatican
203 Venezuela
204 Vietnam
205 **Wallis and Futuna Islands**
206 Western Samoa
207 **Yemen, Republic of**
209 Yugoslavia
210 **Zaire**
211 Zambia
212 Zimbabwe

000 Other (please write below)

10 Which is your first language?

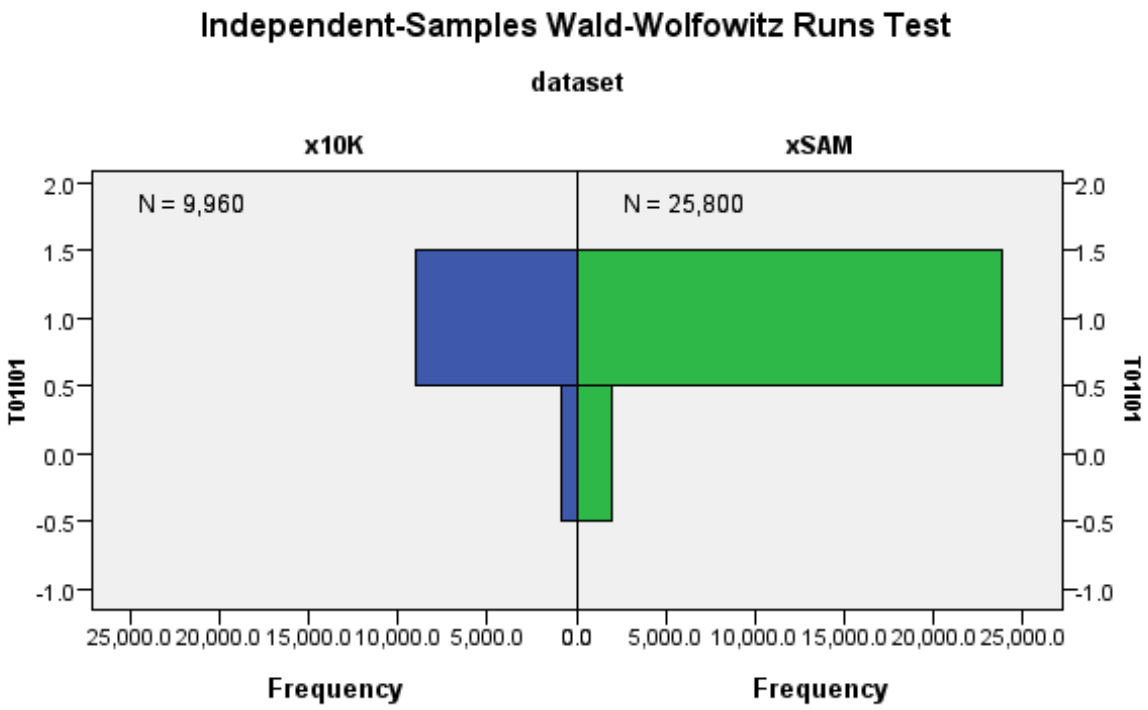
(i.e. your mother tongue).
If it is not listed, please complete the box 'other'.

001 **Afrikaans**
002 Akan
003 Albanian
004 Amharic
005 Arabic
006 Armenian
007 Assamese
008 Aymara
009 Azerbaijani
010 **Baluchi**
011 Bambara
012 Basque
013 Bemba
014 Bengali
015 Bihari
016 Breton
017 Bulgarian
018 Burmese
019 Byelorussian
020 **Catalan**
021 Chinese
022 Croatian
023 Czech
024 **Danish**
025 Dutch
026 **Efik**
027 Estonian
028 Ewe
029 **Faeroese**
030 Farsi
031 Fijian
032 Finnish
033 Flemish
034 French
035 Fulani
036 **Ga**
037 Georgian
038 German
039 Gilbertese
040 Greek
041 Gujarati
042 **Haitian Creole**
043 Hausa
044 Hebrew
045 Hindi
046 Hungarian
047 **Ibo/Igbo**
048 Icelandic
049 Igala
050 Indonesian
051 Italian
052 **Japanese**
053 Javanese
054 **Kannada**
055 Kashmiri
056 Kazakh
057 Khmer
058 Korean
059 **Lao**
060 Latvian
061 Lithuanian
062 Luba
063 Luo
064 Luxemburgish
065 **Malagasy**
066 Malay
067 Malayalam
068 Malinka
069 Maltese
070 Maori
071 Marathi
072 Marshallese
073 Masai
074 Mende
075 Mongolian
076 **Nepali**
077 Norwegian
078 **Oriya**
079 **Palauan**
080 Panjabi
081 Pashto
082 Polish
083 Ponapean
084 Portuguese
085 **Quechua**
086 **Rajasthani**
087 Riff
088 Romanian
089 Romansch
090 Russian
091 **Samoan**
092 Serbian
093 Shona
094 Sindhi
095 Singhalese
096 Slovak
097 Slovene
098 Somali
099 Spanish
100 Swahili
101 Swazi
102 Swedish
129 Swiss German
103 **Tagalog**
104 Tahitian
105 Tamil
106 Tatar
107 Telugu
108 Thai
109 Tibetan
110 Tigrinya
111 Tongan
112 Trukese
113 Tulu
114 Tupi/Guarani
115 Turkish
116 **Uighur**
117 Ukrainian
118 Ulithian
119 Urdu
120 Uzbek
121 **Vietnamese**
122 **Wolof**
123 **Xhosa**
124 **Yao**
125 Yapese
126 Yiddish
127 Yoruba
128 **Zulu**

000 Other (please write below)

This listing of places implies no view regarding questions of sovereignty or status.

Appendix 4: Independent-Samples Wald-Wolfowitz Runs Test results

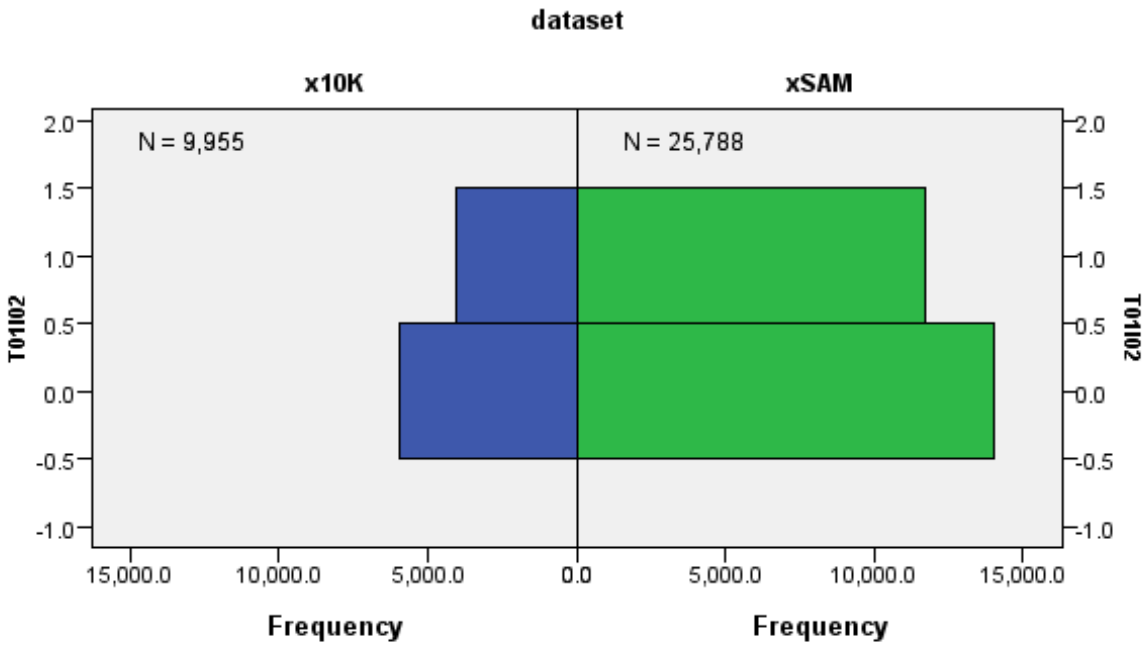


Total N[1]		35,760
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.998
	Standardized Test Statistic[1]	-189.081
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,921.000
	Standard Error[1]	75.998
	Standardized Test Statistic[1]	73.004
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,760 records.

Independent-Samples Wald-Wolfowitz Runs Test

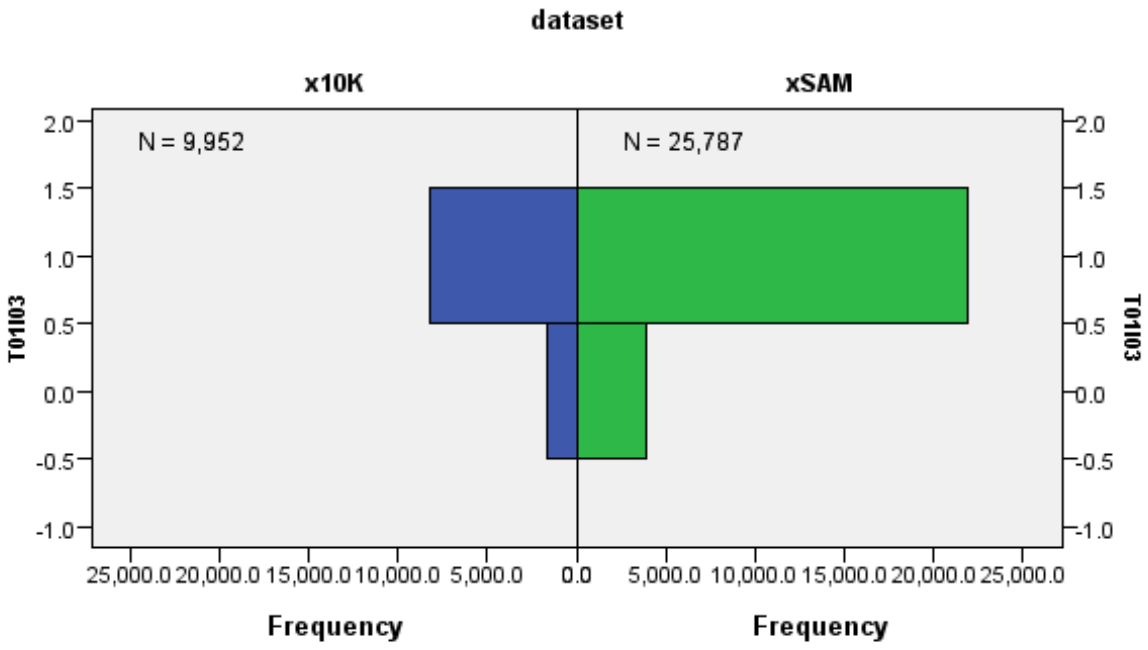


Total N[1]		35,743
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.979
	Standardized Test Statistic[1]	-189.036
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,911.000
	Standard Error[1]	75.979
	Standardized Test Statistic[1]	72.984
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,743 records.

Independent-Samples Wald-Wolfowitz Runs Test

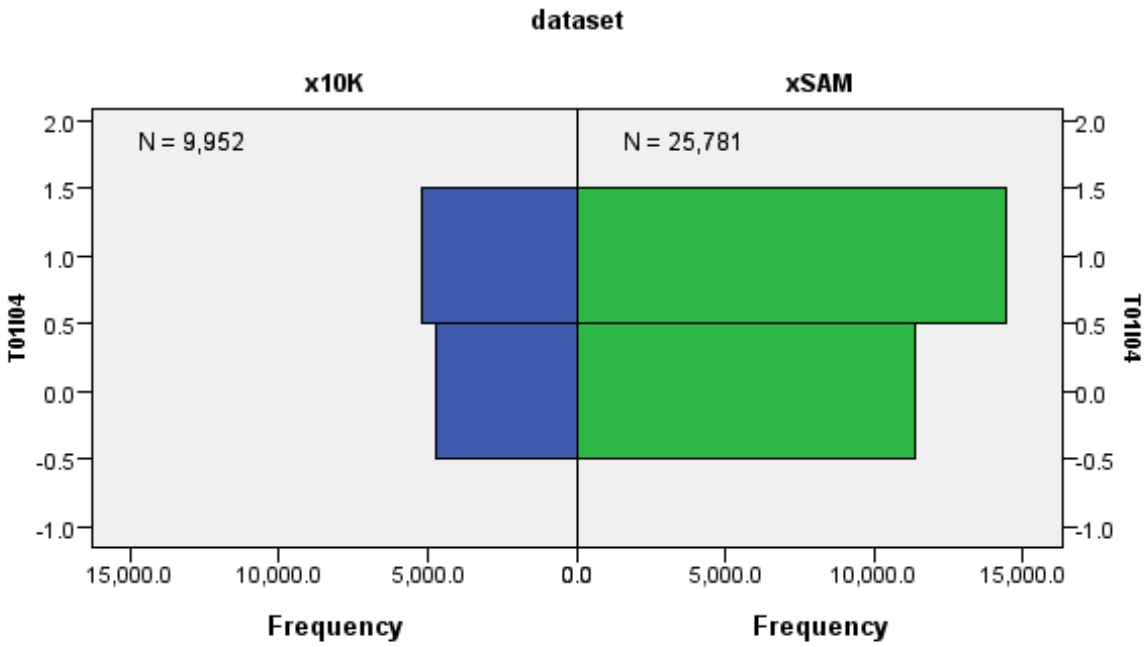


Total N[1]		35,739
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.966
	Standardized Test Statistic[1]	-189.025
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,905.000
	Standard Error[1]	75.966
	Standardized Test Statistic[1]	72.961
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,739 records.

Independent-Samples Wald-Wolfowitz Runs Test

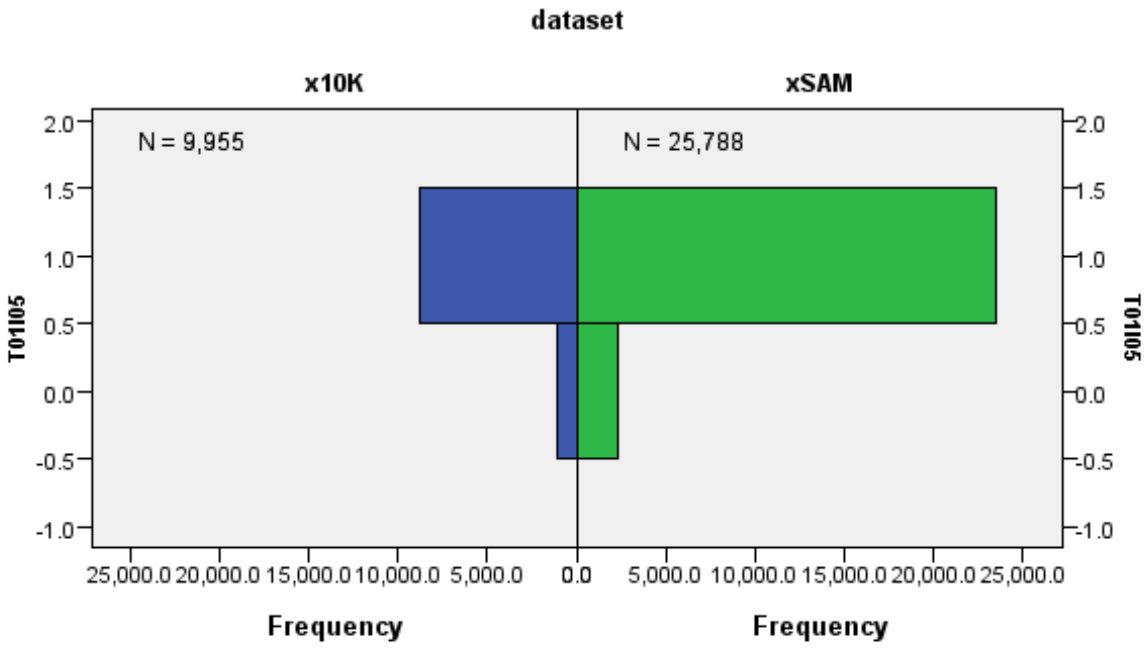


Total N[1]		35,733
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.967
	Standardized Test Statistic[1]	-189.009
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,905.000
	Standard Error[1]	75.967
	Standardized Test Statistic[1]	72.972
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,733 records.

Independent-Samples Wald-Wolfowitz Runs Test

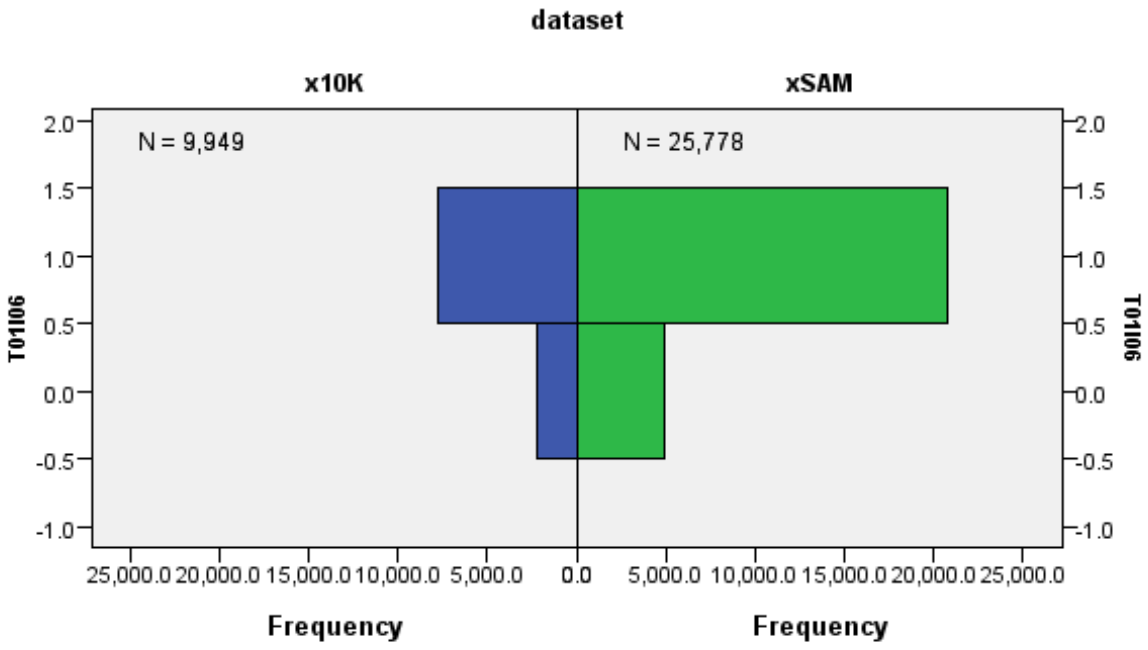


Total N[1]		35,743
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.979
	Standardized Test Statistic[1]	-189.036
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,911.000
	Standard Error[1]	75.979
	Standardized Test Statistic[1]	72.984
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,743 records.

Independent-Samples Wald-Wolfowitz Runs Test

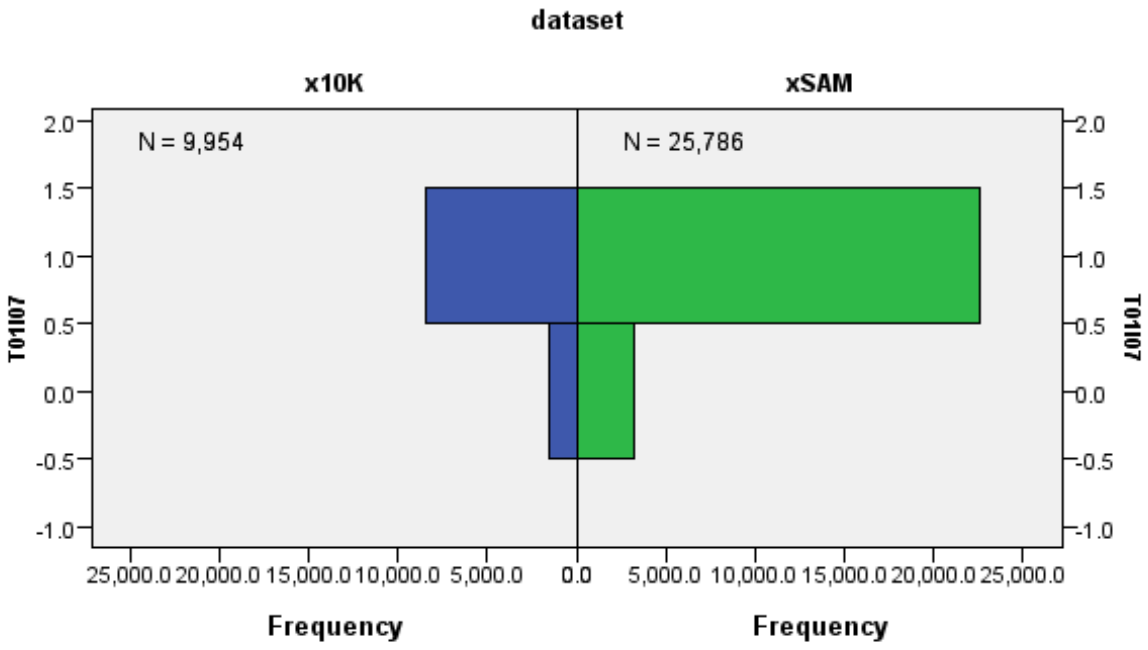


Total N[1]		35,727
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.955
	Standardized Test Statistic[1]	-188.993
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,899.000
	Standard Error[1]	75.955
	Standardized Test Statistic[1]	72.952
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,727 records.

Independent-Samples Wald-Wolfowitz Runs Test

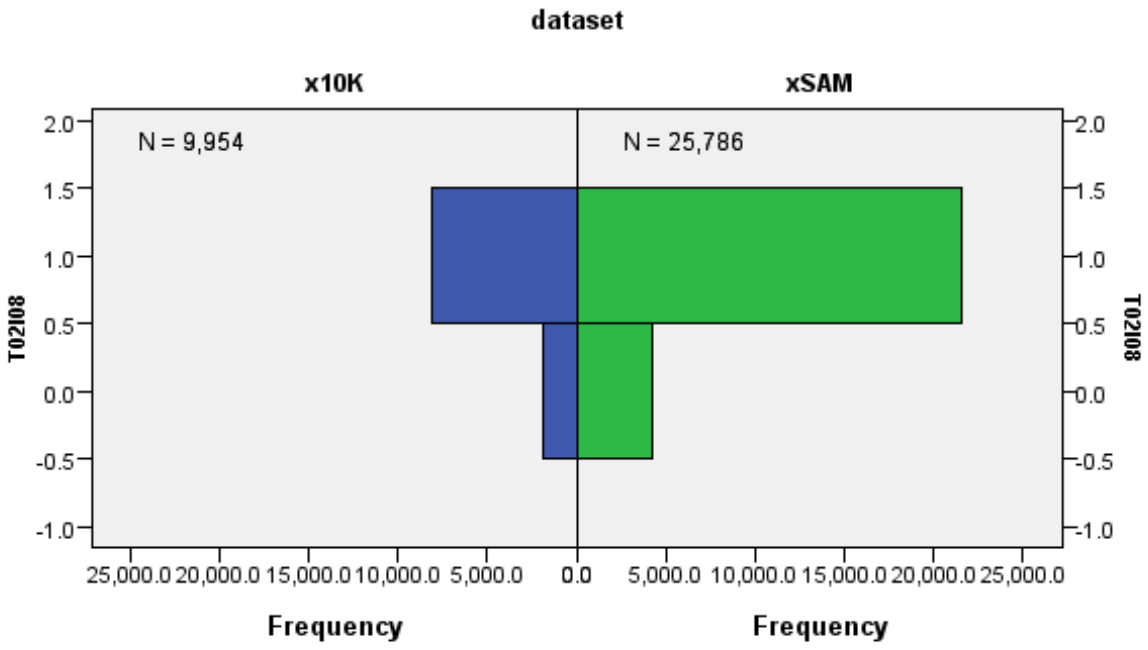


Total N[1]		35,740
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.975
	Standardized Test Statistic[1]	-189.028
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,909.000
	Standard Error[1]	75.975
	Standardized Test Statistic[1]	72.979
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,740 records.

Independent-Samples Wald-Wolfowitz Runs Test

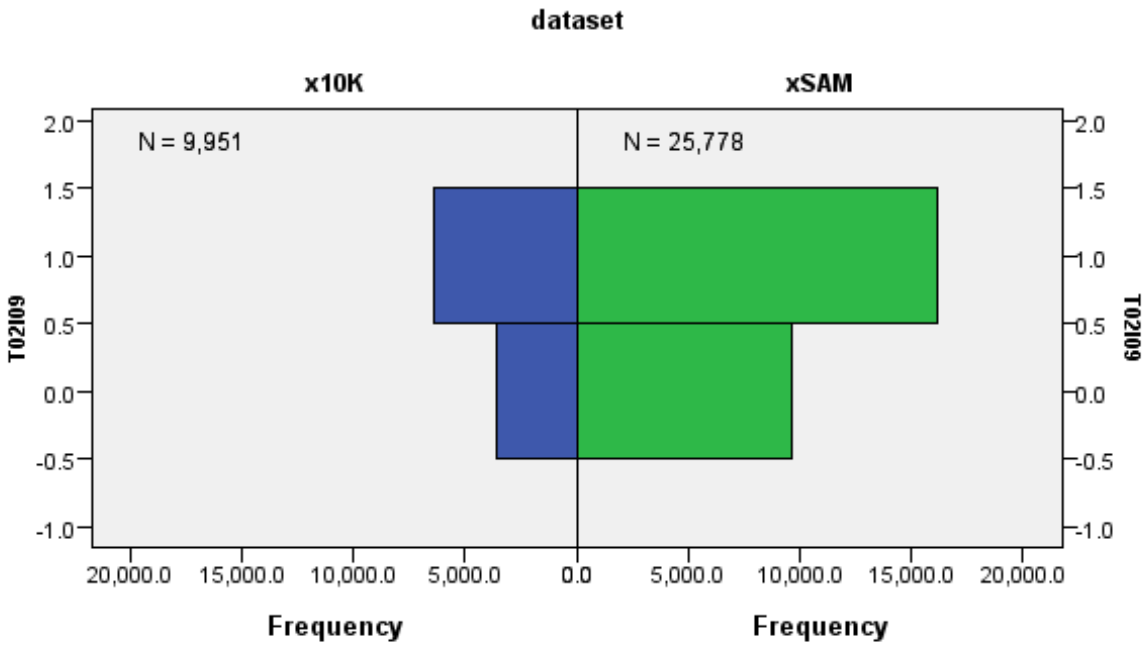


Total N[1]		35,740
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.975
	Standardized Test Statistic[1]	-189.028
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,909.000
	Standard Error[1]	75.975
	Standardized Test Statistic[1]	72.979
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,740 records.

Independent-Samples Wald-Wolfowitz Runs Test

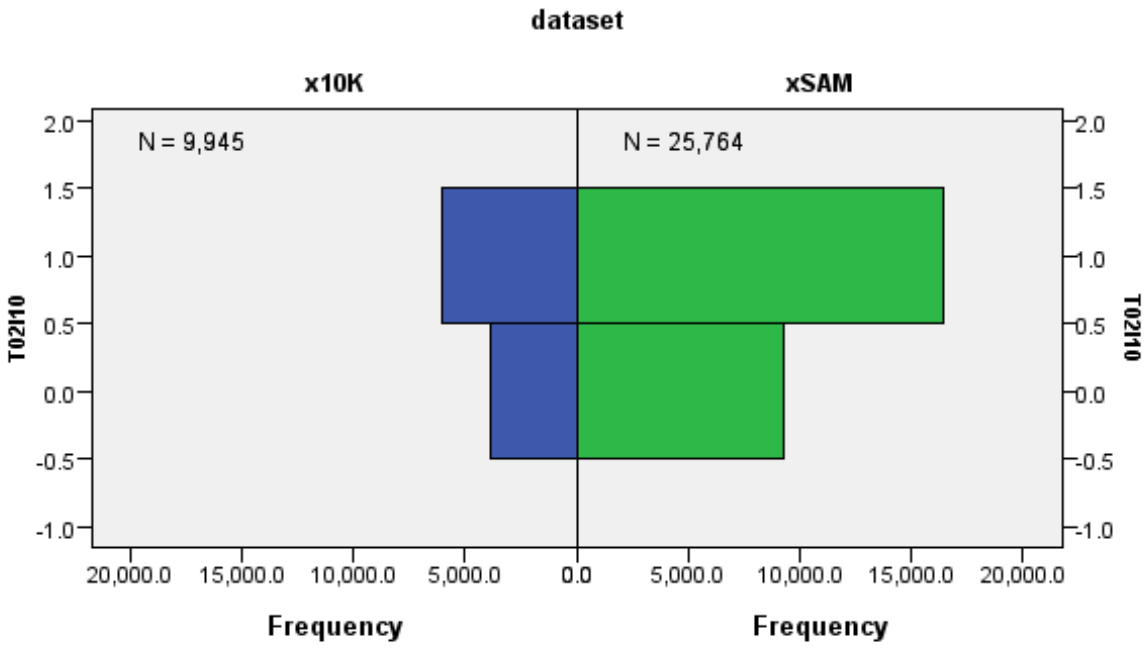


Total N[1]		35,729
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.964
	Standardized Test Statistic[1]	-188.999
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,903.000
	Standard Error[1]	75.964
	Standardized Test Statistic[1]	72.969
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,729 records.

Independent-Samples Wald-Wolfowitz Runs Test

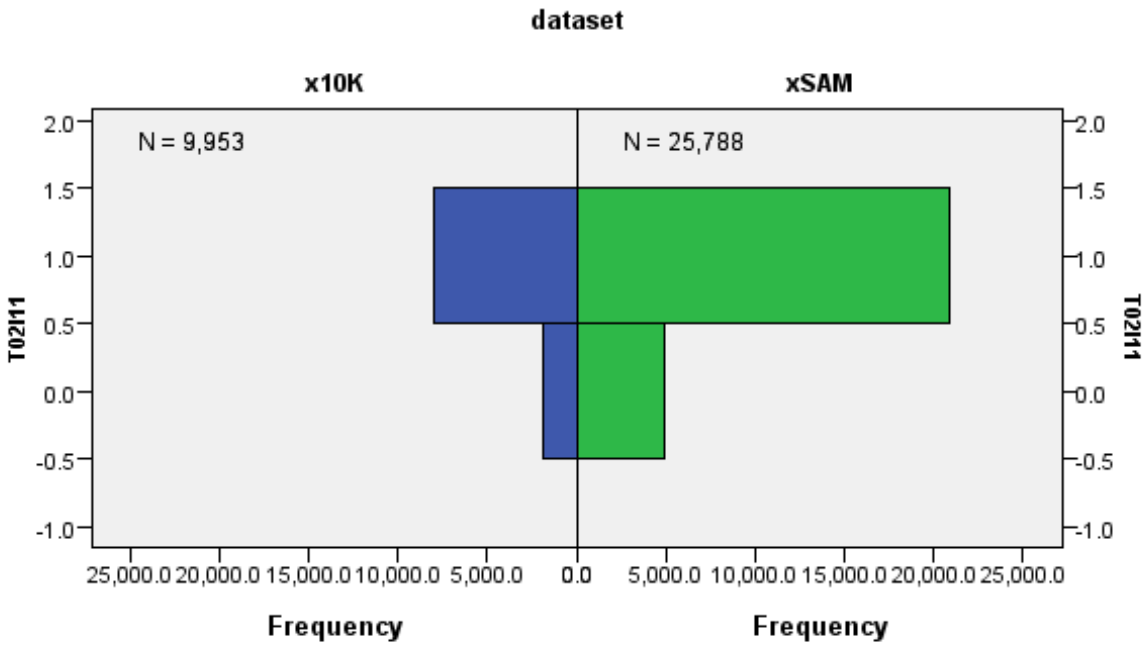


Total N[1]		35,709
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.940
	Standardized Test Statistic[1]	-188.946
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,891.000
	Standard Error[1]	75.940
	Standardized Test Statistic[1]	72.944
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,709 records.

Independent-Samples Wald-Wolfowitz Runs Test

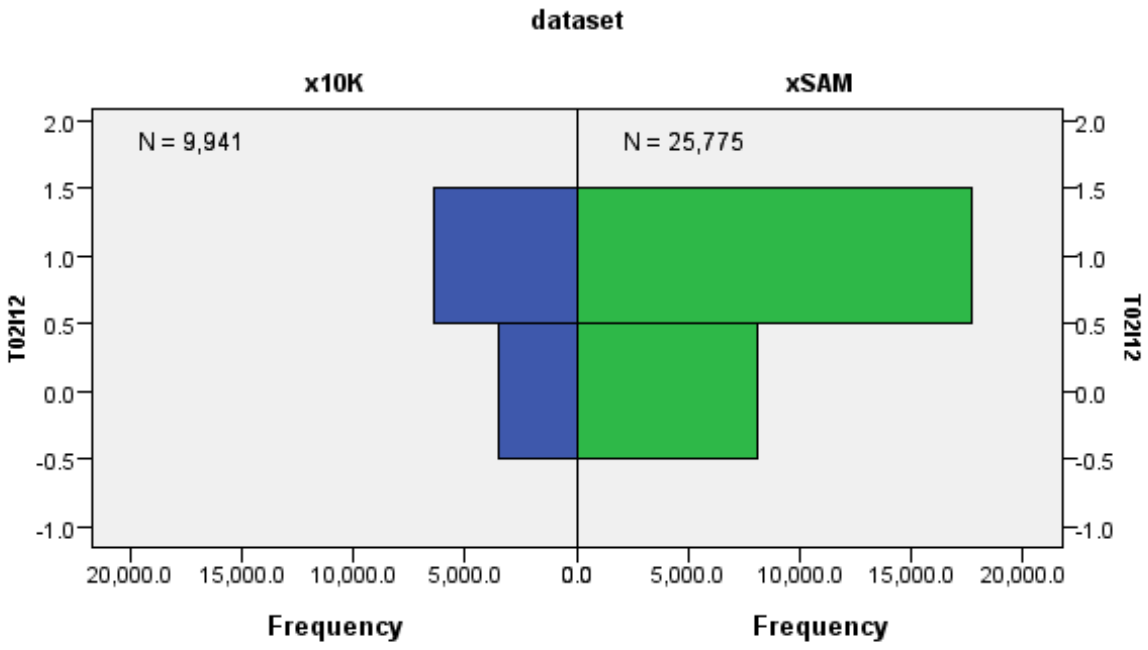


Total N[1]		35,741
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.970
	Standardized Test Statistic[1]	-189.031
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,907.000
	Standard Error[1]	75.970
	Standardized Test Statistic[1]	72.967
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,741 records.

Independent-Samples Wald-Wolfowitz Runs Test

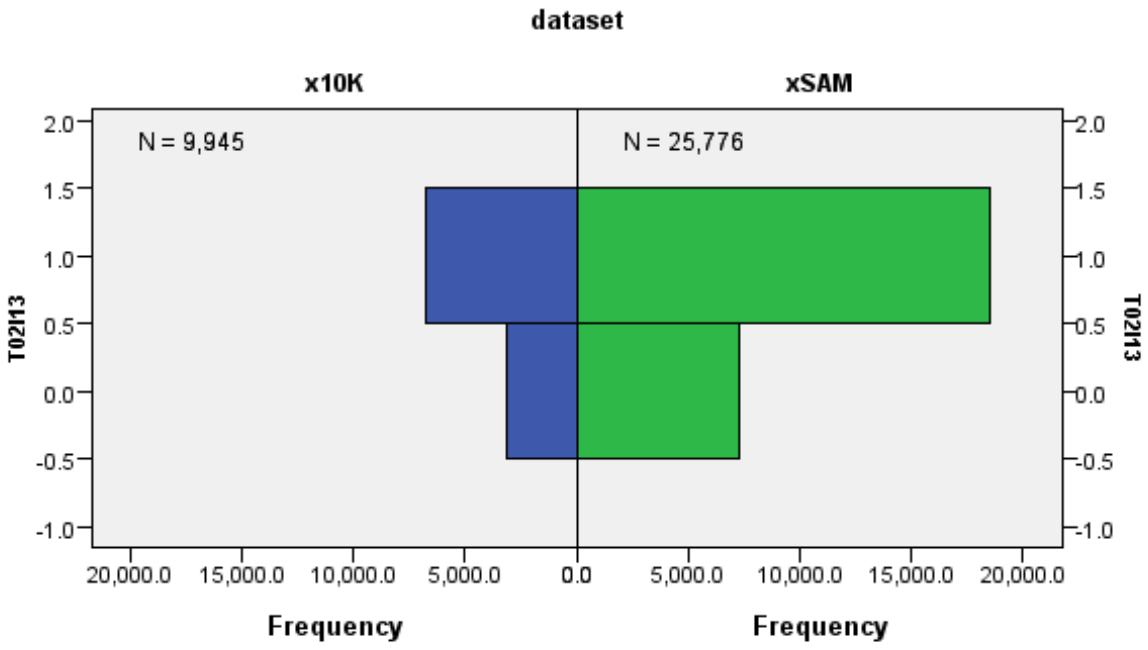


Total N[1]		35,716
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.920
	Standardized Test Statistic[1]	-188.964
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,883.000
	Standard Error[1]	75.920
	Standardized Test Statistic[1]	72.891
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,716 records.

Independent-Samples Wald-Wolfowitz Runs Test

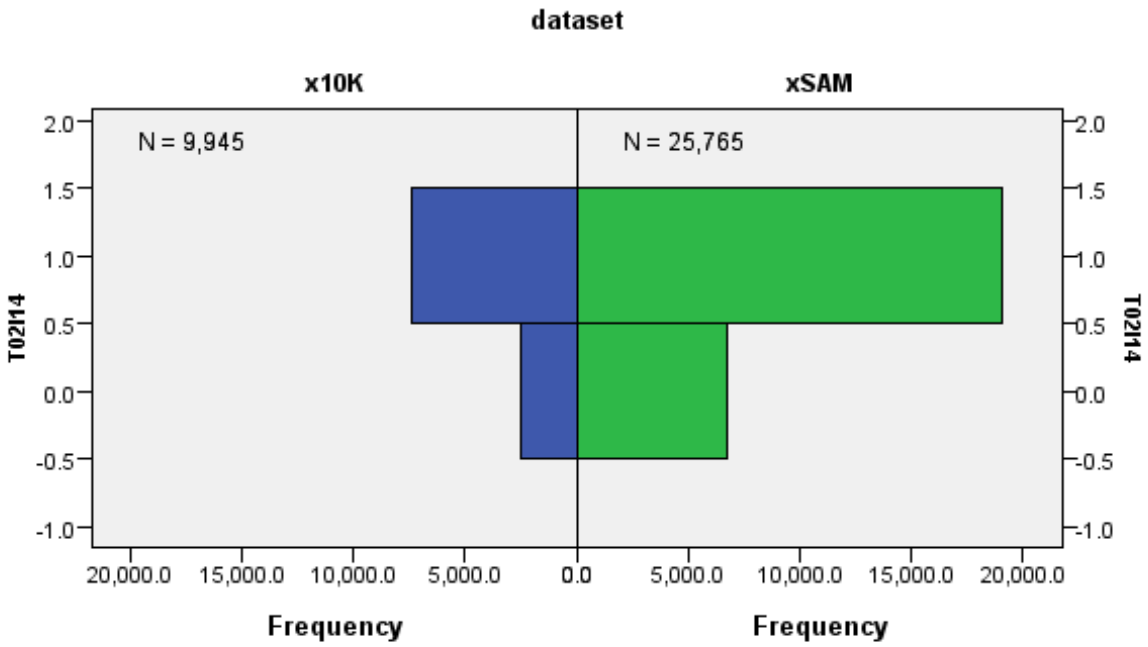


Total N[1]		35,721
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.937
	Standardized Test Statistic[1]	-188.978
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,891.000
	Standard Error[1]	75.937
	Standardized Test Statistic[1]	72.922
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,721 records.

Independent-Samples Wald-Wolfowitz Runs Test

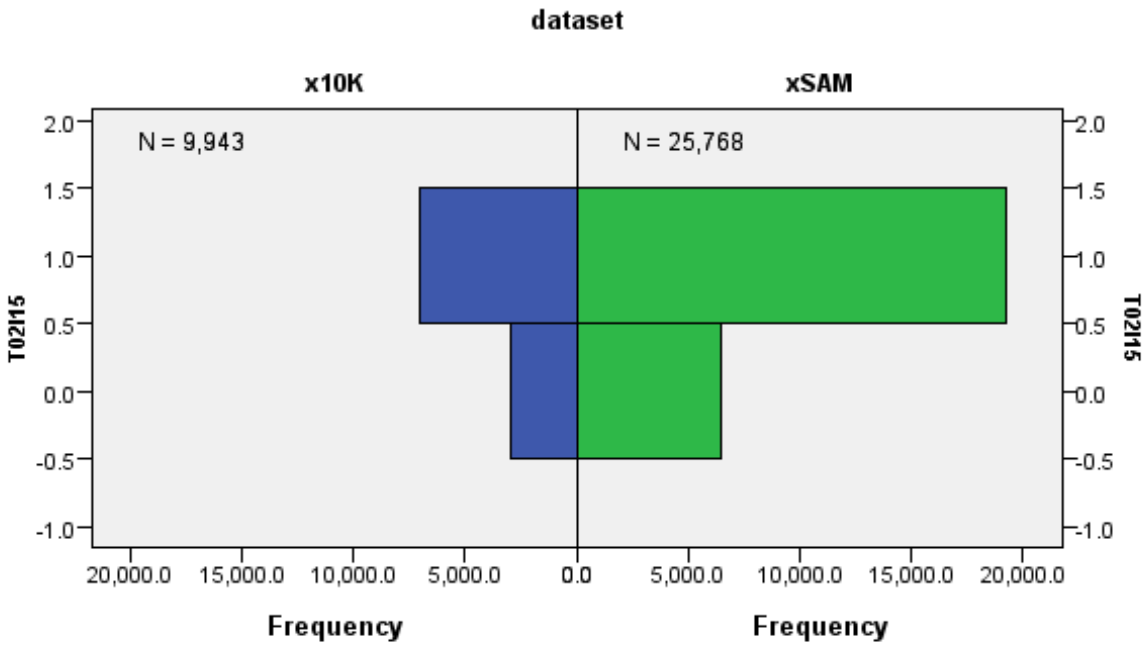


Total N[1]		35,710
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.940
	Standardized Test Statistic[1]	-188.948
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,891.000
	Standard Error[1]	75.940
	Standardized Test Statistic[1]	72.942
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,710 records.

Independent-Samples Wald-Wolfowitz Runs Test

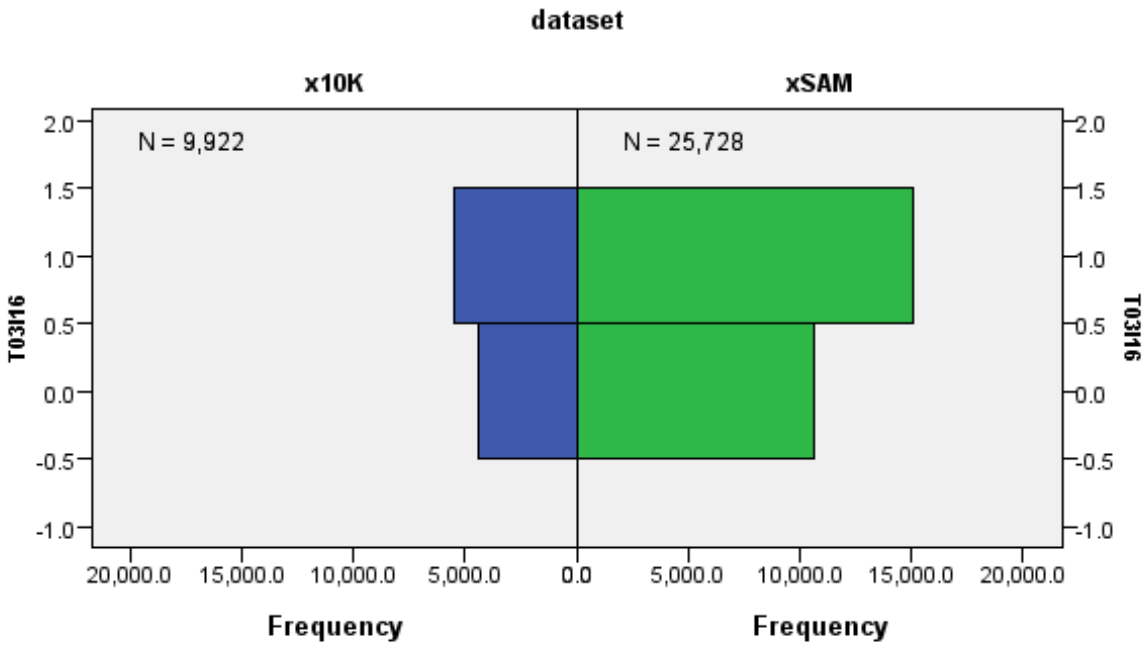


Total N[1]		35,711
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.930
	Standardized Test Statistic[1]	-188.951
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,887.000
	Standard Error[1]	75.930
	Standardized Test Statistic[1]	72.920
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,711 records.

Independent-Samples Wald-Wolfowitz Runs Test

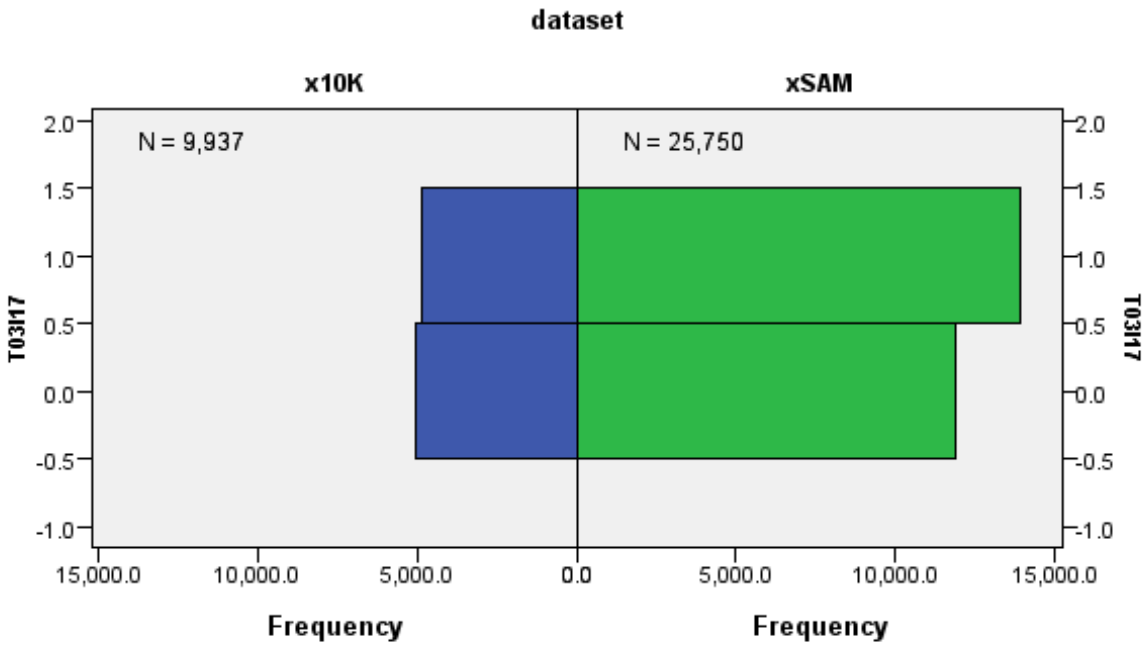


Total N[1]		35,650
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.847
	Standardized Test Statistic[1]	-188.790
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,845.000
	Standard Error[1]	75.847
	Standardized Test Statistic[1]	72.817
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,650 records.

Independent-Samples Wald-Wolfowitz Runs Test

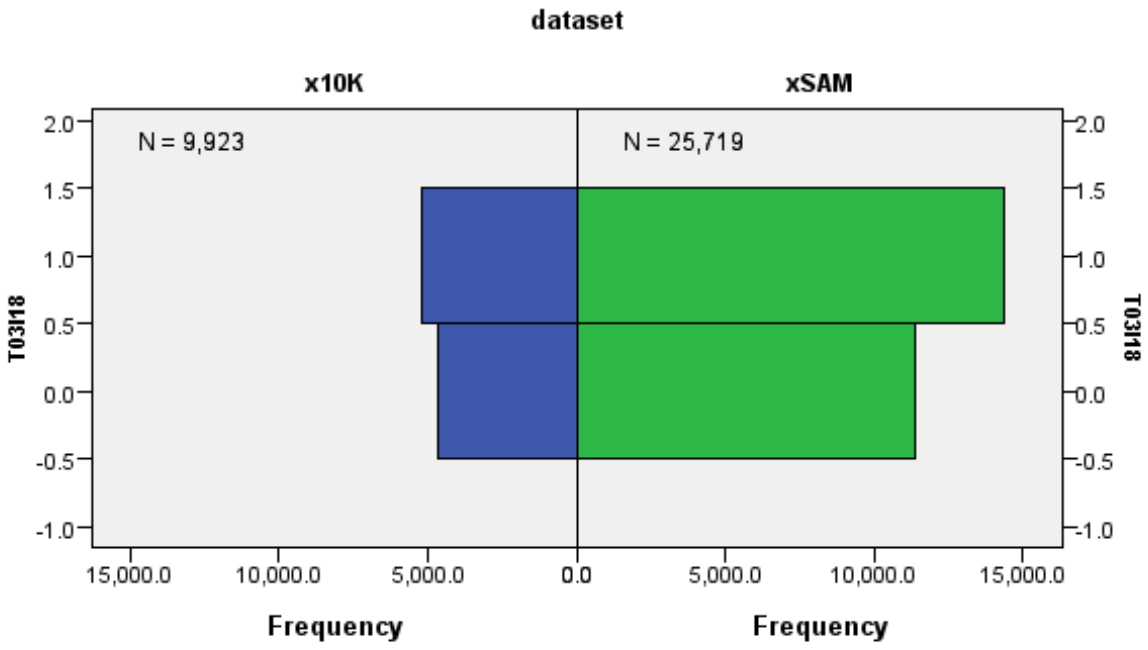


Total N[1]		35,687
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.908
	Standardized Test Statistic[1]	-188.888
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,875.000
	Standard Error[1]	75.908
	Standardized Test Statistic[1]	72.902
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,687 records.

Independent-Samples Wald-Wolfowitz Runs Test

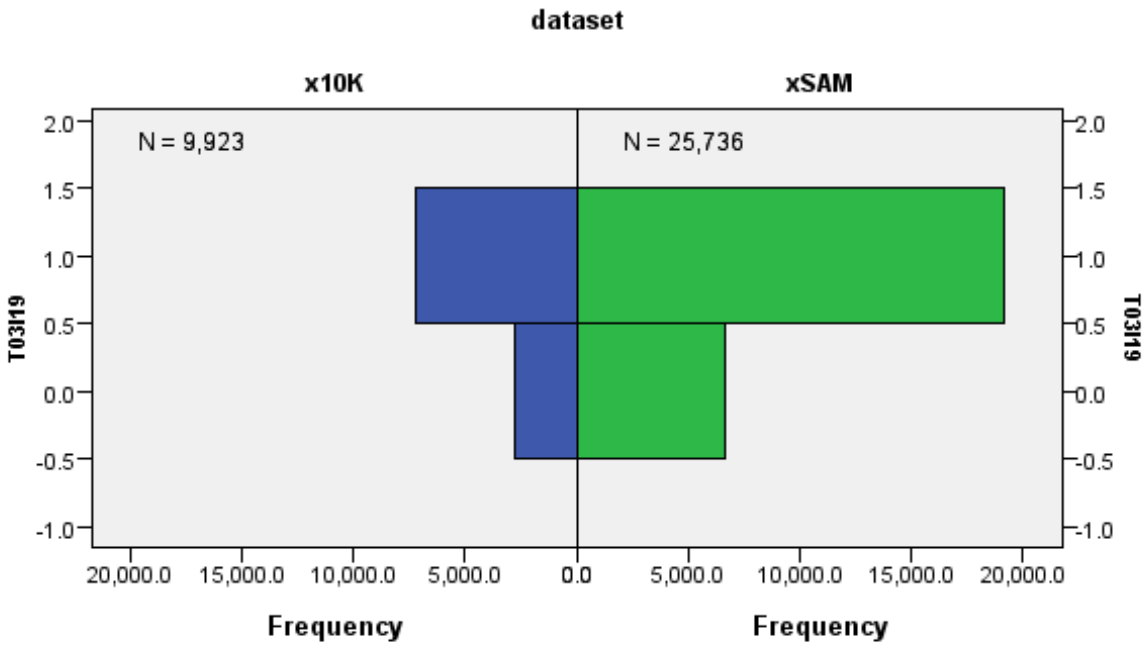


Total N[1]		35,642
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.853
	Standardized Test Statistic[1]	-188.768
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,847.000
	Standard Error[1]	75.853
	Standardized Test Statistic[1]	72.842
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,642 records.

Independent-Samples Wald-Wolfowitz Runs Test

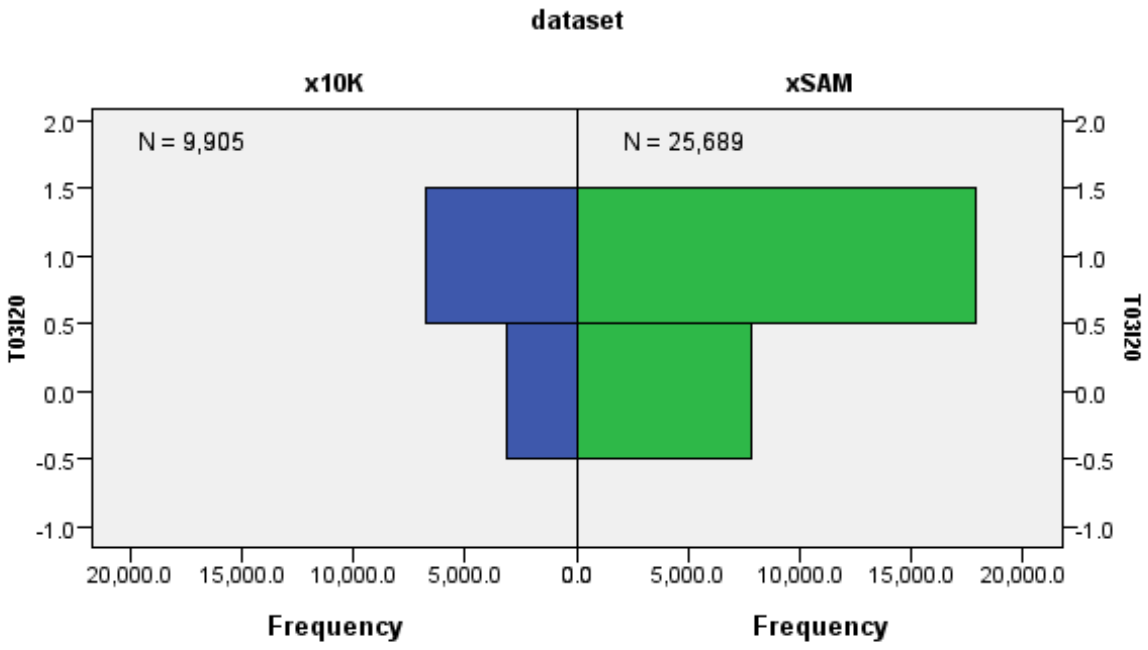


Total N[1]		35,659
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.849
	Standardized Test Statistic[1]	-188.813
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,847.000
	Standard Error[1]	75.849
	Standardized Test Statistic[1]	72.811
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,659 records.

Independent-Samples Wald-Wolfowitz Runs Test

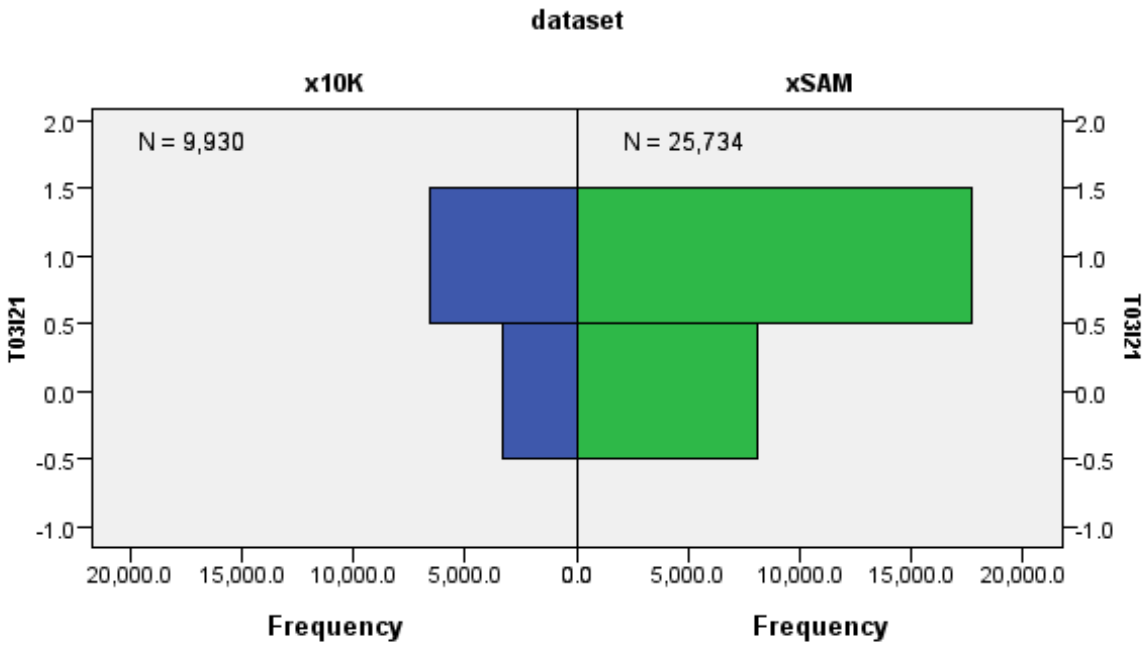


Total N[1]		35,594
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.780
	Standardized Test Statistic[1]	-188.641
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,811.000
	Standard Error[1]	75.780
	Standardized Test Statistic[1]	72.745
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,594 records.

Independent-Samples Wald-Wolfowitz Runs Test

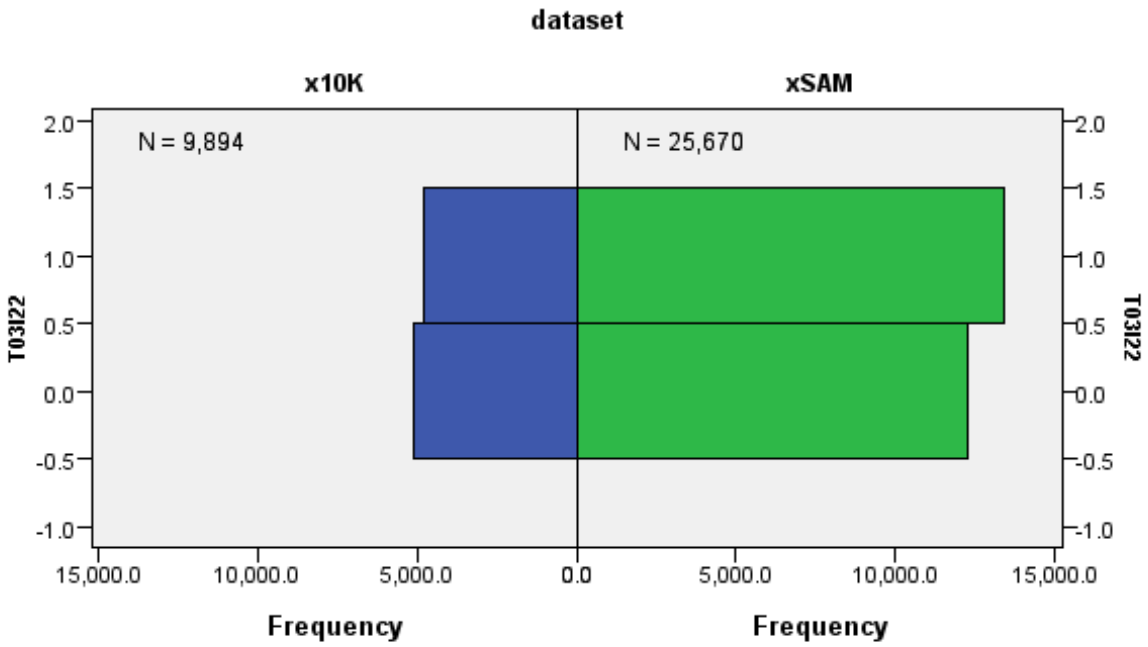


Total N[1]		35,664
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.881
	Standardized Test Statistic[1]	-188.827
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,861.000
	Standard Error[1]	75.881
	Standardized Test Statistic[1]	72.873
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,664 records.

Independent-Samples Wald-Wolfowitz Runs Test

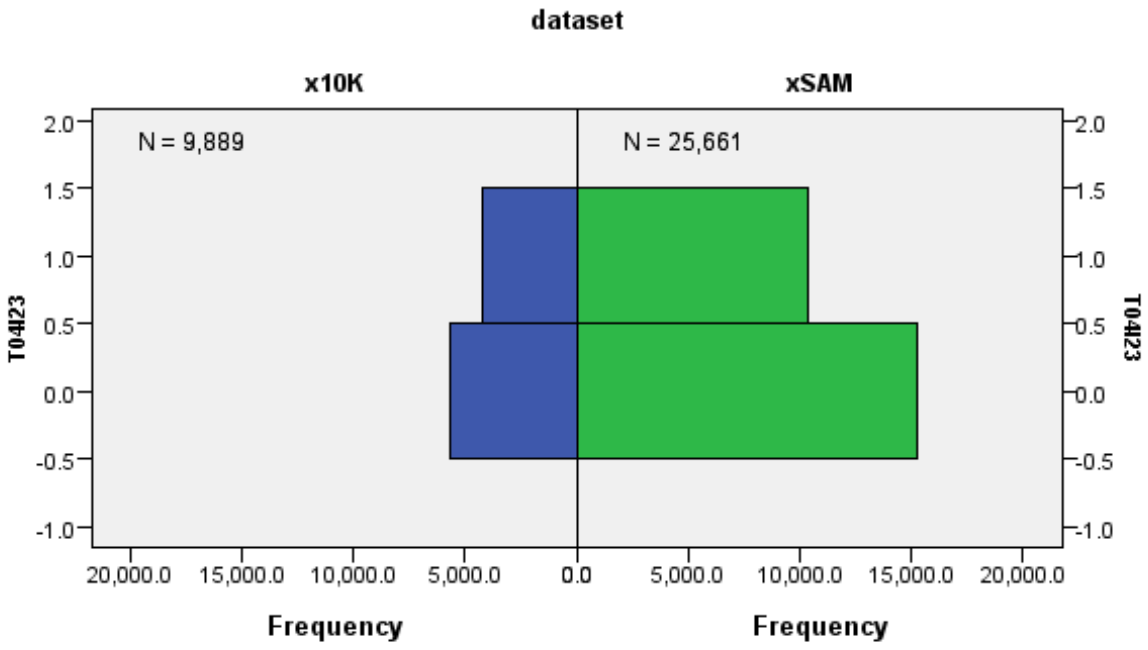


Total N[1]		35,564
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.736
	Standardized Test Statistic[1]	-188.562
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,789.000
	Standard Error[1]	75.736
	Standardized Test Statistic[1]	72.688
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,564 records.

Independent-Samples Wald-Wolfowitz Runs Test

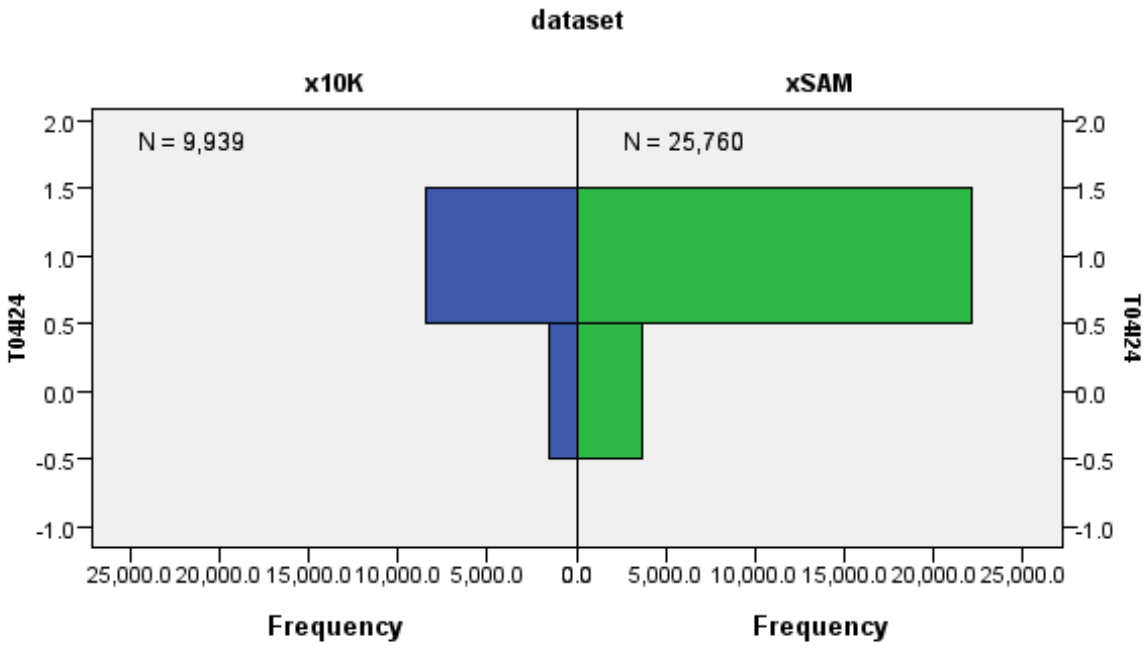


Total N[1]		35,550
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.716
	Standardized Test Statistic[1]	-188.525
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,779.000
	Standard Error[1]	75.716
	Standardized Test Statistic[1]	72.662
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,550 records.

Independent-Samples Wald-Wolfowitz Runs Test

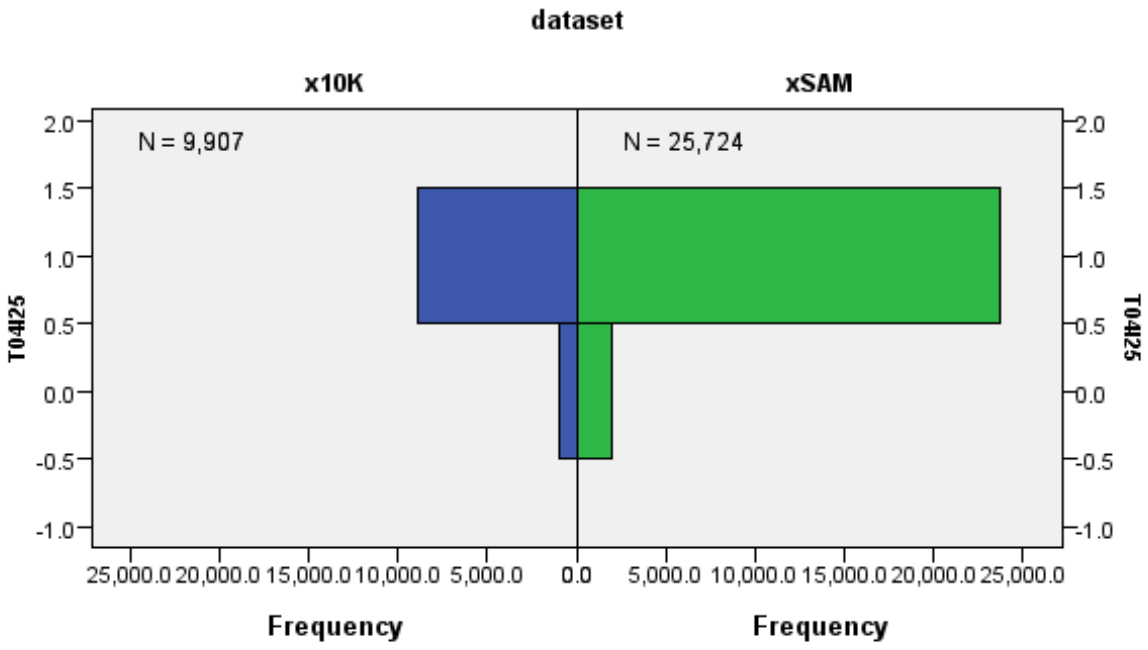


Total N[1]		35,699
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.915
	Standardized Test Statistic[1]	-188.919
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,879.000
	Standard Error[1]	75.915
	Standardized Test Statistic[1]	72.901
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,699 records.

Independent-Samples Wald-Wolfowitz Runs Test

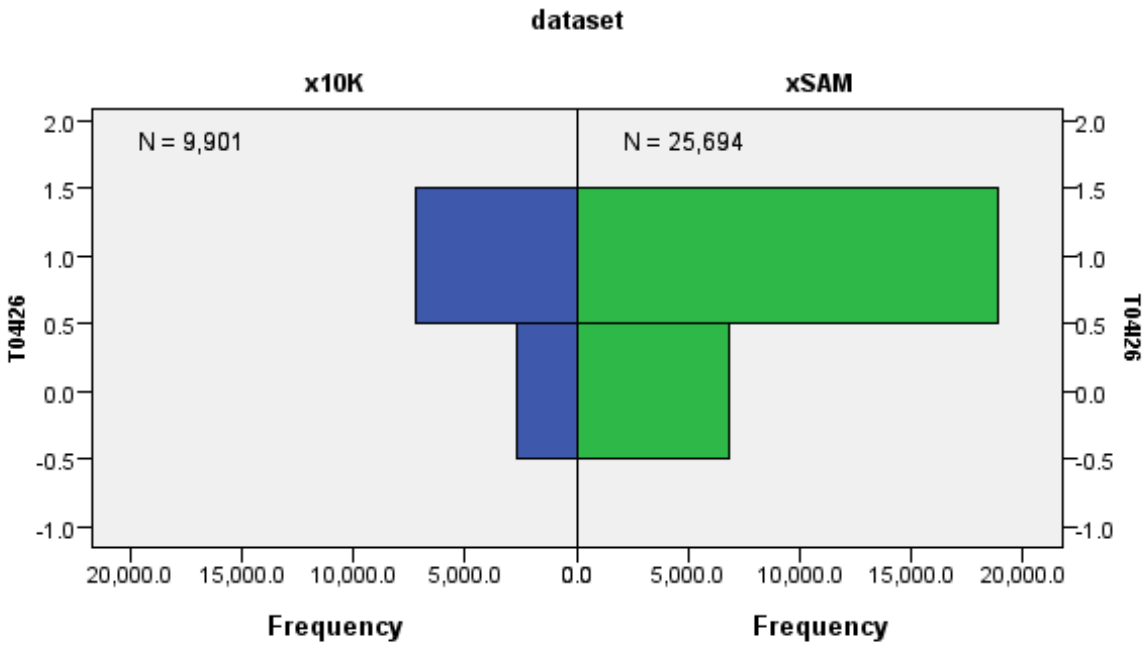


Total N[1]		35,631
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.781
	Standardized Test Statistic[1]	-188.739
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,815.000
	Standard Error[1]	75.781
	Standardized Test Statistic[1]	72.699
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,631 records.

Independent-Samples Wald-Wolfowitz Runs Test

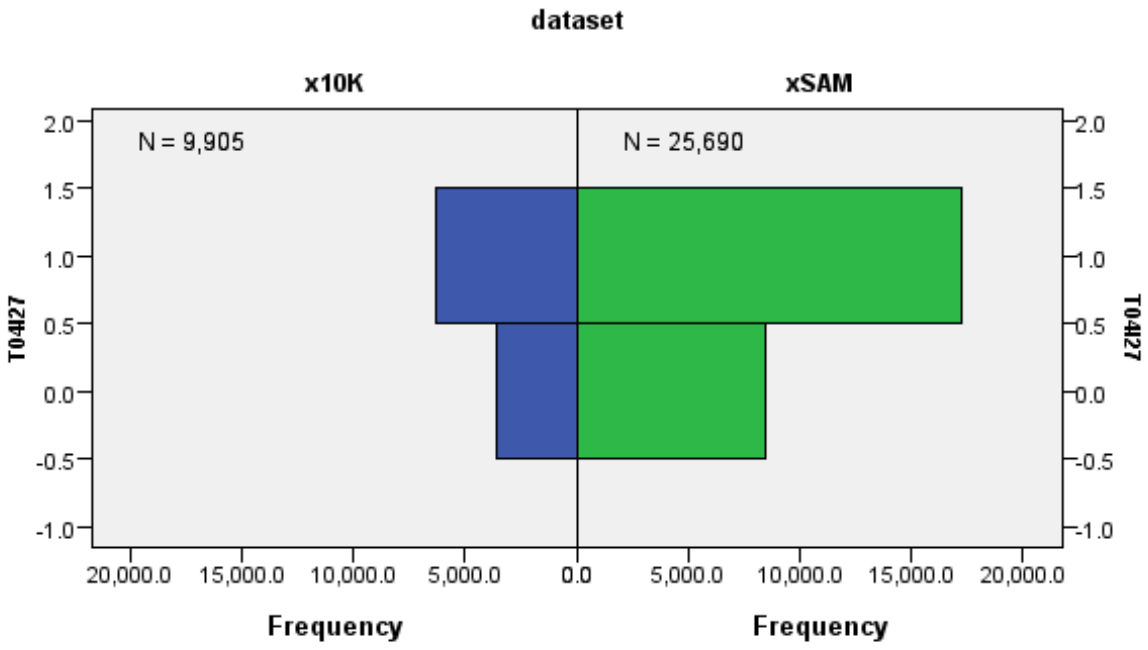


Total N[1]		35,595
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.761
	Standardized Test Statistic[1]	-188.644
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,803.000
	Standard Error[1]	75.761
	Standardized Test Statistic[1]	72.703
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,595 records.

Independent-Samples Wald-Wolfowitz Runs Test

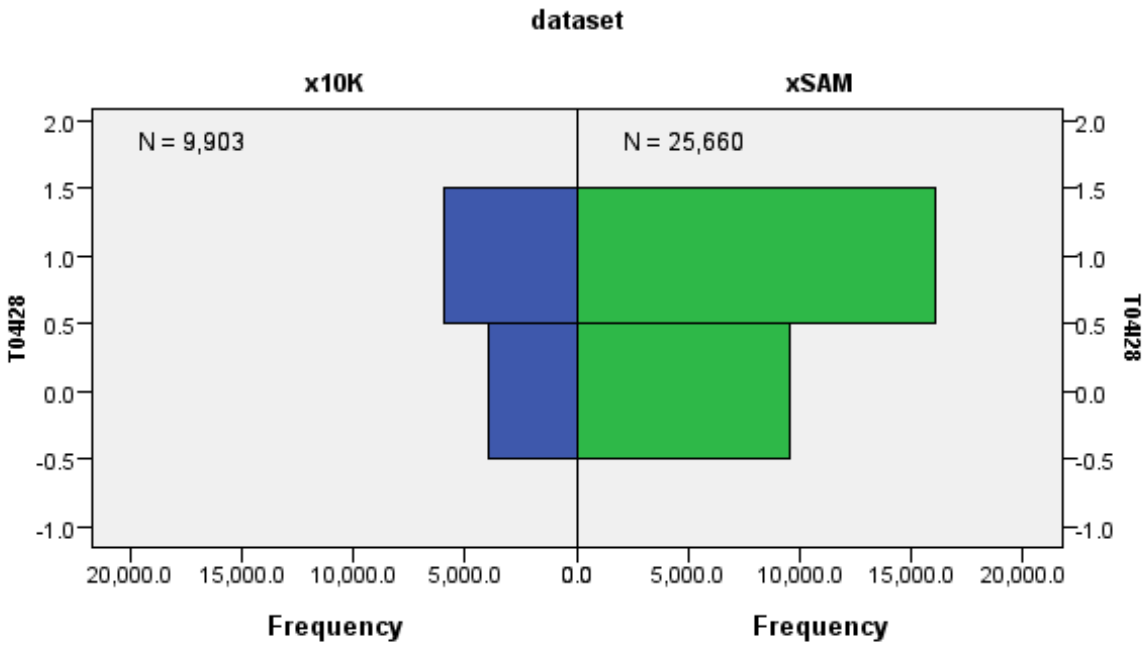


Total N[1]		35,595
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.780
	Standardized Test Statistic[1]	-188.644
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,811.000
	Standard Error[1]	75.780
	Standardized Test Statistic[1]	72.743
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,595 records.

Independent-Samples Wald-Wolfowitz Runs Test

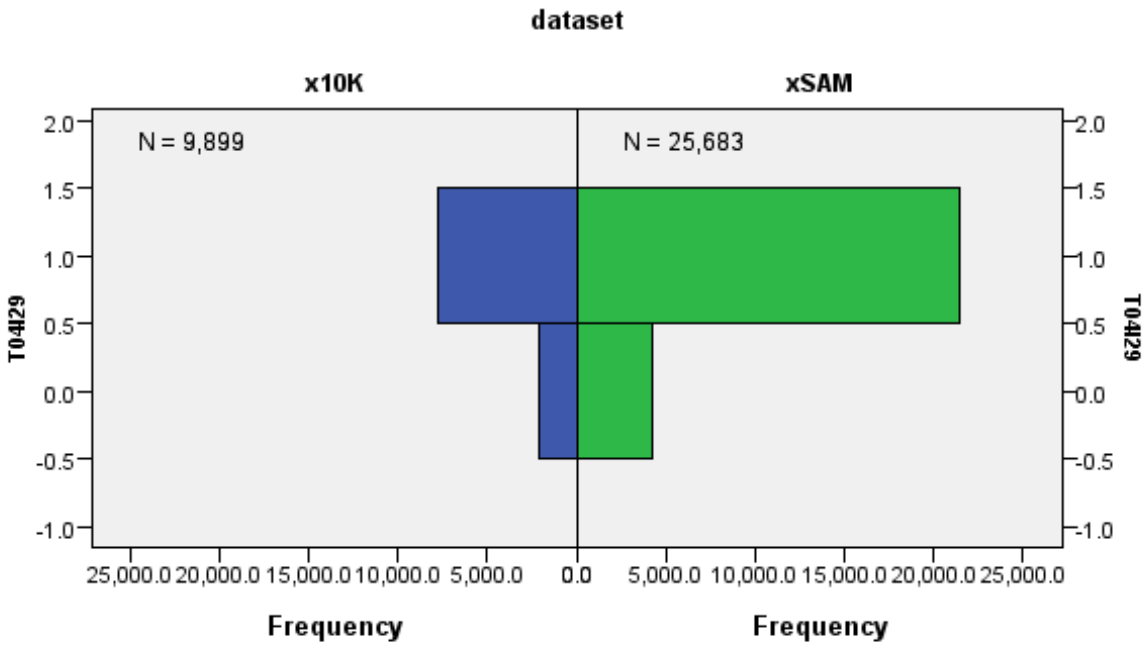


Total N[1]		35,563
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.779
	Standardized Test Statistic[1]	-188.559
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,807.000
	Standard Error[1]	75.779
	Standardized Test Statistic[1]	72.781
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,563 records.

Independent-Samples Wald-Wolfowitz Runs Test

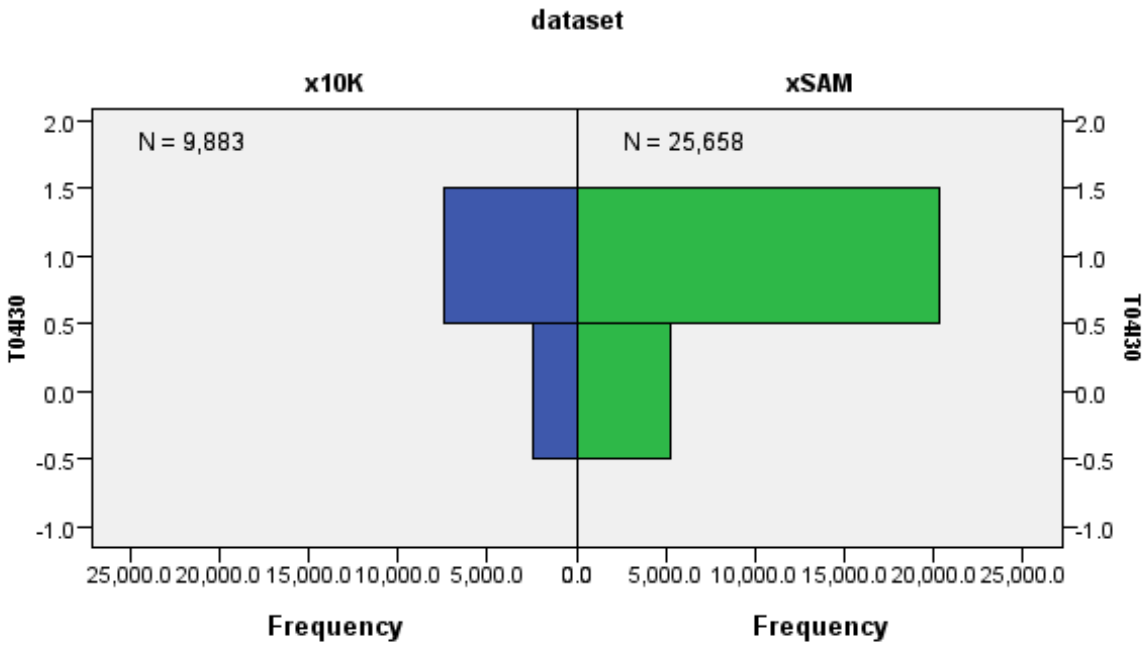


Total N[1]		35,582
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.755
	Standardized Test Statistic[1]	-188.609
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,799.000
	Standard Error[1]	75.755
	Standardized Test Statistic[1]	72.706
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,582 records.

Independent-Samples Wald-Wolfowitz Runs Test

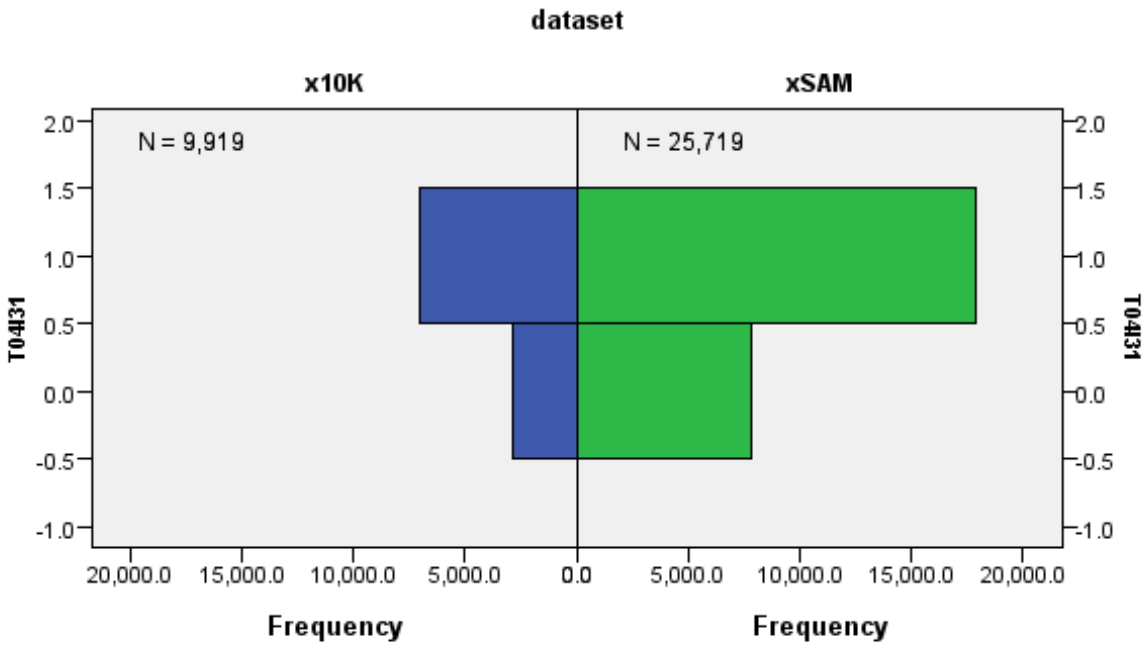


Total N[1]		35,541
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.690
	Standardized Test Statistic[1]	-188.501
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,767.000
	Standard Error[1]	75.690
	Standardized Test Statistic[1]	72.617
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,541 records.

Independent-Samples Wald-Wolfowitz Runs Test

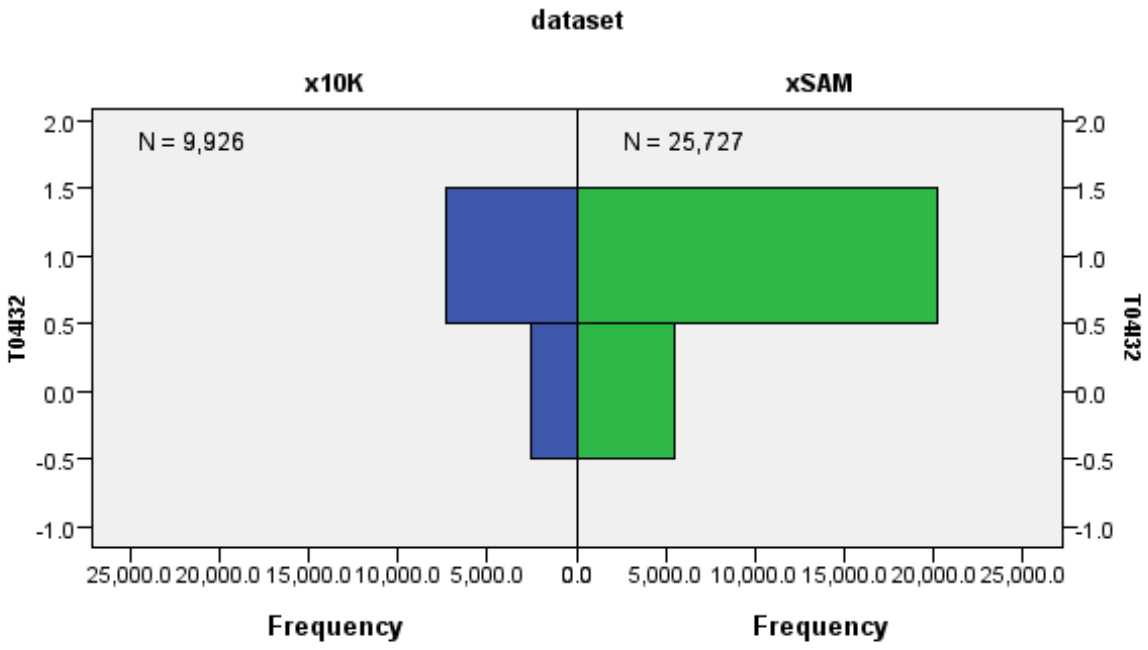


Total N[1]		35,638
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.836
	Standardized Test Statistic[1]	-188.758
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,839.000
	Standard Error[1]	75.836
	Standardized Test Statistic[1]	72.808
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,638 records.

Independent-Samples Wald-Wolfowitz Runs Test

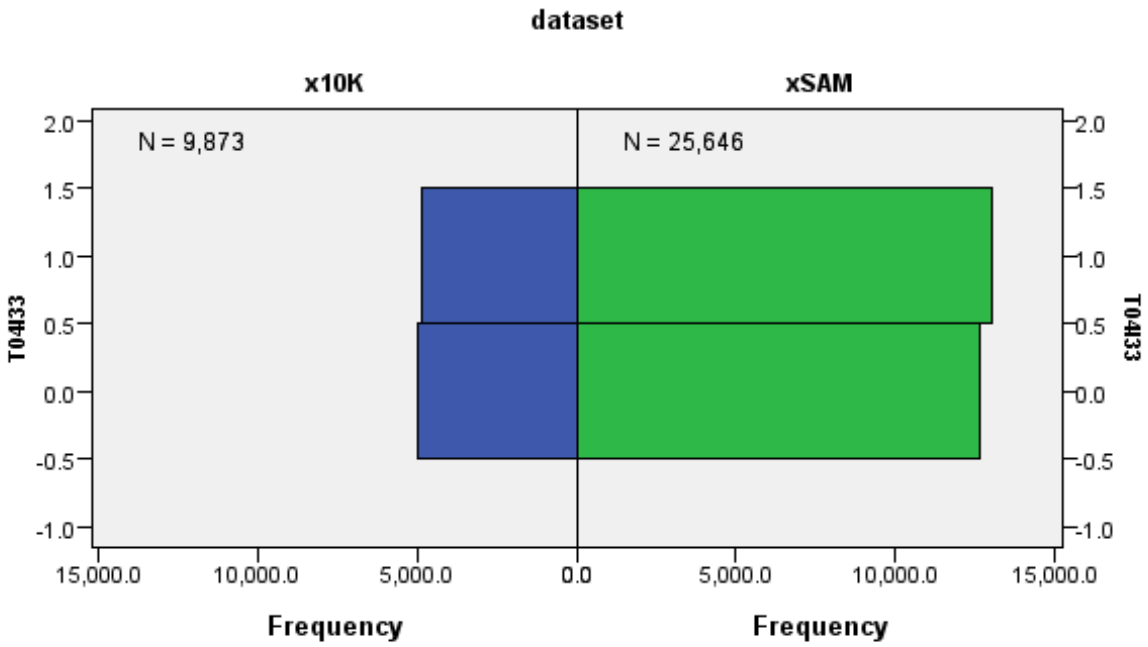


Total N[1]		35,653
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.865
	Standardized Test Statistic[1]	-188.798
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,853.000
	Standard Error[1]	75.865
	Standardized Test Statistic[1]	72.852
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,653 records.

Independent-Samples Wald-Wolfowitz Runs Test

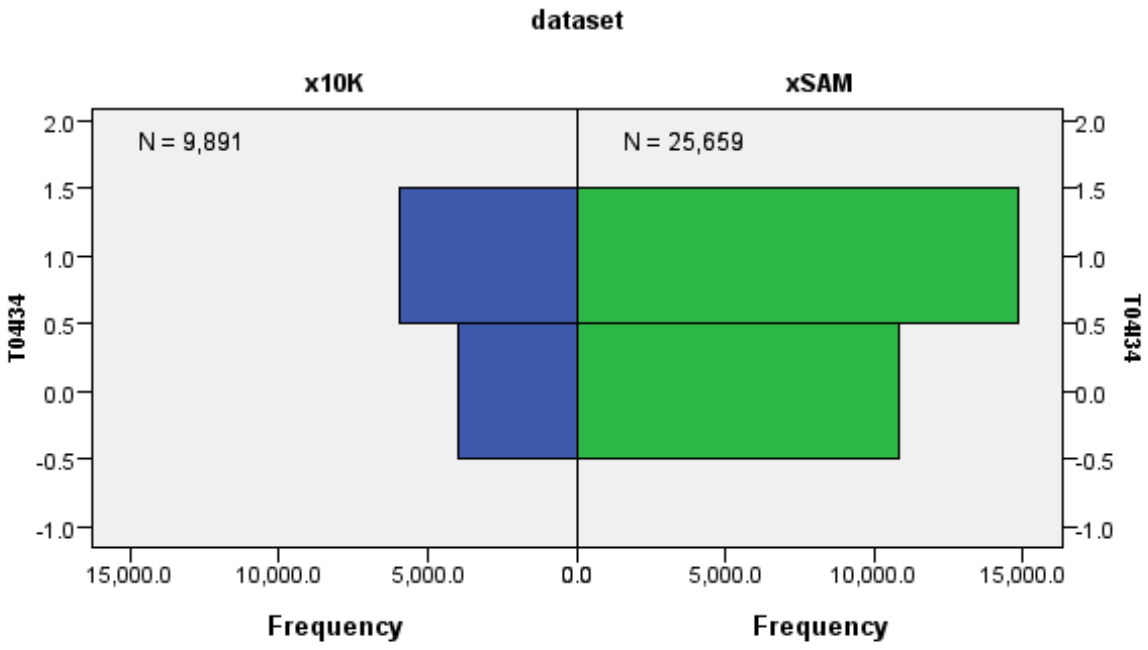


Total N[1]		35,519
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.648
	Standardized Test Statistic[1]	-188.442
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,747.000
	Standard Error[1]	75.648
	Standardized Test Statistic[1]	72.555
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,519 records.

Independent-Samples Wald-Wolfowitz Runs Test

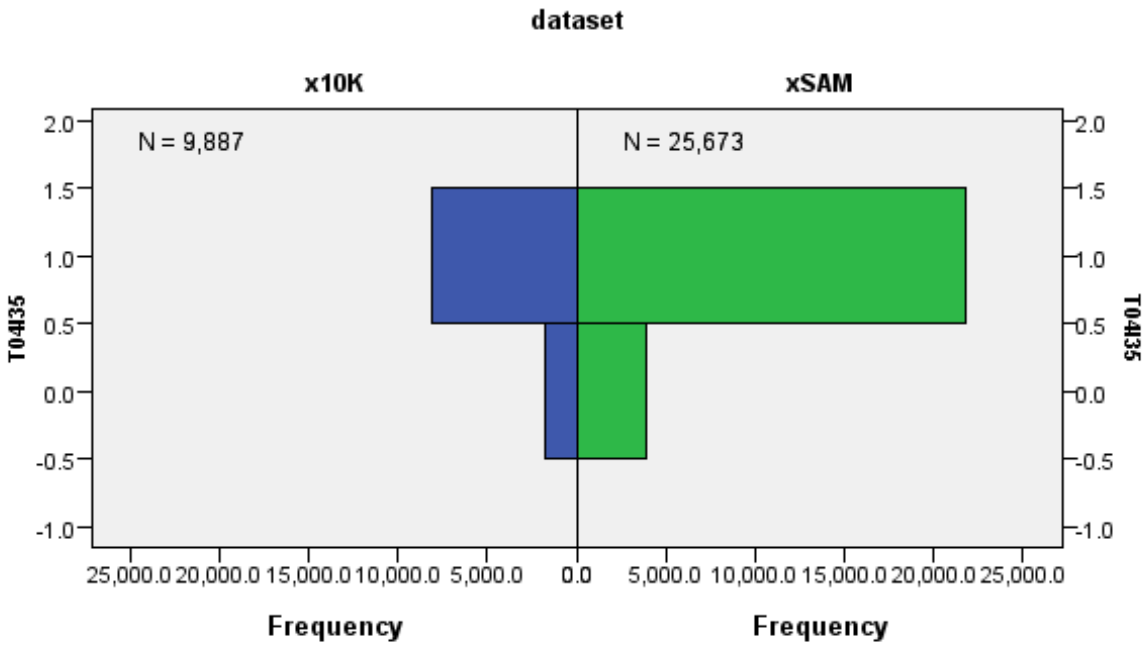


Total N[1]		35,550
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.725
	Standardized Test Statistic[1]	-188.525
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,783.000
	Standard Error[1]	75.725
	Standardized Test Statistic[1]	72.682
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,550 records.

Independent-Samples Wald-Wolfowitz Runs Test



Total N[1]		35,560
Minimum Possible	**Test Statistic**[1]	3.000
	Standard Error[1]	75.704
	Standardized Test Statistic[1]	-188.551
	Asymptotic Sig. (2-sided test)[1]	.000
Maximum Possible	**Test Statistic**[1]	19,775.000
	Standard Error[1]	75.704
	Standardized Test Statistic[1]	72.624
	Asymptotic Sig. (2-sided test)[1]	1.000

[1]The test statistic is the number of runs.

1. There are 2 inter-group ties involving 35,560 records.

Appendix 5: summary of response matrices

Table 75 Summary of response data for Part 1, crossed, cleaned data set

	T01I01	T01I02	T01I03	T01I04	T01I05	T01I06	T01I07
A	0.14%	0.48%	0.50%	0.97%	91.00%	2.11%	0.41%
B	0.65%	45.49%	3.15%	22.81%	3.41%	4.10%	2.30%
C	0.17%	1.64%	1.83%	2.77%	0.51%	3.87%	87.49%
D	0.97%	5.94%	1.81%	55.94%	0.93%	4.66%	2.89%
E	92.36%	1.15%	2.26%	1.27%	0.15%	0.67%	0.32%
F	0.63%	42.53%	3.48%	7.84%	0.62%	2.69%	2.07%
G	0.18%	1.05%	2.00%	4.60%	3.10%	80.66%	3.98%
H	4.89%	1.66%	84.90%	3.70%	0.22%	1.15%	0.47%
O	0.01%	0.06%	0.06%	0.09%	0.06%	0.10%	0.07%

Table 76 Summary of response data for Part 1, sample data set

	T01I01	T01I02	T01I03	T01I04	T01I05	T01I06	T01I07
A	0.27%	0.69%	0.75%	1.46%	88.65%	2.58%	0.71%
B	0.72%	40.43%	3.40%	23.81%	4.03%	4.93%	2.69%
C	0.28%	2.01%	2.29%	3.05%	0.72%	4.74%	84.85%
D	1.33%	6.04%	2.18%	52.48%	1.49%	4.73%	3.49%
E	91.00%	1.45%	2.30%	1.47%	0.21%	0.86%	0.44%
F	0.80%	45.99%	4.19%	8.37%	0.70%	2.66%	2.16%
G	0.27%	1.37%	2.03%	4.74%	3.83%	78.02%	4.92%
H	5.31%	1.97%	82.77%	4.53%	0.31%	1.36%	0.66%
O	0.01%	0.06%	0.09%	0.09%	0.06%	0.12%	0.07%

Table 77 Summary of response data for Part 2, crossed, cleaned data set

	T02I08	T02I09	T02I10	T02I11	T02I12	T02I13	T02I14	T02I15
A	12.28%	62.54%	13.27%	80.90%	15.51%	4.77%	12.11%	5.27%
B	3.26%	11.77%	63.83%	16.48%	4.01%	71.58%	3.48%	74.65%
C	0.81%	20.81%	11.37%	1.46%	68.51%	19.57%	10.53%	10.37%
D	83.58%	4.78%	11.39%	1.09%	11.87%	3.97%	73.74%	9.58%
O	0.07%	0.10%	0.15%	0.06%	0.11%	0.10%	0.15%	0.14%

Table 78 Summary of response data for Part 2, sample data set

	T02I08	T02I09	T02I10	T02I11	T02I12	T02I13	T02I14	T02I15
A	13.71%	64.05%	15.24%	80.53%	15.63%	5.02%	11.70%	5.49%
B	3.79%	13.09%	60.91%	16.82%	4.15%	68.09%	3.49%	70.23%
C	1.31%	17.44%	12.26%	1.55%	64.30%	22.01%	10.12%	12.51%
D	81.12%	5.32%	11.43%	1.02%	15.72%	4.73%	74.53%	11.59%
O	0.07%	0.10%	0.16%	0.08%	0.20%	0.16%	0.16%	0.18%

Table 79 Summary of response data for Part 3, crossed, cleaned data set

	T03I16	T03I17	T03I18	T03I19	T03I20	T03I21	T03I22
A	1.91%	35.87%	55.63%	0.74%	0.35%	0.54%	1.91%
B	18.03%	0.66%	1.69%	2.04%	69.34%	2.56%	3.65%
C	58.45%	1.47%	3.00%	3.37%	12.30%	5.15%	13.72%
D	3.54%	6.31%	16.70%	5.01%	1.92%	4.78%	5.60%
E	1.95%	0.87%	3.23%	74.12%	1.24%	2.50%	2.80%
F	2.53%	0.38%	11.91%	9.85%	4.05%	13.65%	51.97%
G	4.04%	0.46%	3.32%	4.06%	4.76%	68.53%	11.63%
H	9.25%	53.78%	4.20%	0.54%	5.60%	2.02%	8.19%
O	0.29%	0.21%	0.33%	0.26%	0.44%	0.27%	0.52%

Table 80 Summary of response data for Part 3, sample data set

	T03I16	T03I17	T03I18	T03I19	T03I20	T03I21	T03I22
A	2.17%	39.32%	52.54%	0.69%	0.39%	0.54%	1.66%
B	17.74%	0.82%	1.78%	2.44%	67.69%	2.64%	4.42%
C	55.56%	1.55%	3.19%	3.75%	13.01%	5.06%	15.06%
D	3.88%	7.01%	17.99%	5.93%	2.20%	4.79%	5.91%
E	1.92%	0.96%	3.47%	72.09%	1.42%	2.31%	2.59%
F	2.74%	0.38%	12.18%	10.22%	4.25%	15.75%	47.85%
G	4.60%	0.66%	3.87%	3.86%	4.67%	66.20%	12.41%
H	11.01%	49.05%	4.60%	0.63%	5.81%	2.40%	9.44%
O	0.39%	0.24%	0.38%	0.38%	0.56%	0.31%	0.67%

Table 81 Summary of response data for Part 4, crossed, cleaned data set

	T04I23	T04I24	T04I25	T04I26	T04I27	T04I28	T04I29	T04I30	T04I31	T04I32	T04I33	T04I34	T04I35
A	6.05%	85.63%	0.86%	73.22%	1.79%	62.33%	1.96%	4.53%	69.29%	3.92%	10.63%	30.38%	3.82%
B	42.98%	6.94%	2.77%	3.95%	66.78%	6.72%	83.06%	6.00%	1.43%	15.22%	50.39%	4.20%	5.85%
C	40.25%	5.15%	92.07%	18.79%	5.20%	13.20%	8.75%	10.01%	2.69%	78.31%	6.00%	7.31%	84.22%
D	10.17%	2.11%	3.99%	3.62%	25.80%	17.20%	5.76%	78.89%	26.27%	2.26%	32.36%	57.55%	5.61%
O	0.55%	0.17%	0.31%	0.42%	0.44%	0.55%	0.46%	0.56%	0.33%	0.29%	0.61%	0.56%	0.50%

Table 82 Summary of response data for Part 4, sample data set

	T04I23	T04I24	T04I25	T04I26	T04I27	T04I28	T04I29	T04I30	T04I31	T04I32	T04I33	T04I34	T04I35
A	5.53%	84.66%	1.09%	72.04%	2.01%	59.36%	2.36%	5.45%	70.38%	5.60%	10.97%	27.60%	4.66%
B	41.76%	7.66%	4.42%	4.76%	63.39%	7.77%	77.87%	7.02%	1.53%	17.14%	48.83%	4.42%	7.34%
C	42.64%	4.94%	89.39%	18.57%	6.07%	14.30%	11.30%	11.82%	2.81%	73.77%	7.14%	7.65%	81.36%
D	9.35%	2.52%	4.56%	4.03%	27.97%	17.99%	7.84%	74.93%	24.86%	3.14%	32.18%	59.63%	5.90%
O	0.72%	0.22%	0.54%	0.60%	0.56%	0.58%	0.62%	0.78%	0.42%	0.35%	0.88%	0.70%	0.74%

Appendix 6: summary of the score matrices

Table 83 Summary of score data for Part 1, crossed, cleaned data set

Score	T01I01	T01I02	T01I03	T01I04	T01I05	T01I06	T01I07
0	7.63%	54.45%	15.04%	43.98%	8.94%	19.25%	12.44%
1	92.36%	45.49%	84.90%	55.94%	91.00%	80.66%	87.49%
O	0.01%	0.06%	0.06%	0.09%	0.06%	0.10%	0.07%

Table 84 Summary of score data for Part 1, sample data set

Score	T01I01	T01I02	T01I03	T01I04	T01I05	T01I06	T01I07
0	8.99%	59.51%	17.14%	47.42%	11.29%	21.86%	15.08%
1	91.00%	40.43%	82.77%	52.48%	88.65%	78.02%	84.85%
O	0.01%	0.06%	0.09%	0.09%	0.06%	0.12%	0.07%

Table 85 Summary of score data for Part 2, crossed, cleaned data set

Score	T02I08	T02I09	T02I10	T02I11	T02I12	T02I13	T02I14	T02I15
0	16.35%	37.36%	36.02%	19.04%	31.38%	28.31%	26.12%	25.22%
1	83.58%	62.54%	63.83%	80.90%	68.51%	71.58%	73.74%	74.65%
O	0.07%	0.10%	0.15%	0.06%	0.11%	0.10%	0.15%	0.14%

Table 86 Summary of score data for Part 2, sample data set

Score	T02I08	T02I09	T02I10	T02I11	T02I12	T02I13	T02I14	T02I15
0	18.81%	35.85%	38.93%	19.39%	35.50%	31.75%	25.31%	29.59%
1	81.12%	64.05%	60.91%	80.53%	64.30%	68.09%	74.53%	70.23%
O	0.07%	0.10%	0.16%	0.08%	0.20%	0.16%	0.16%	0.18%

Table 87 Summary of score data for Part 3, crossed, cleaned data set

Score	T03I16	T03I17	T03I18	T03I19	T03I20	T03I21	T03I22
0	41.25%	46.01%	44.05%	25.62%	30.22%	31.20%	47.52%
1	58.45%	53.78%	55.63%	74.12%	69.34%	68.53%	51.97%
O	0.29%	0.21%	0.33%	0.26%	0.44%	0.27%	0.52%

Table 88 Summary of score data for Part 3, sample data set

Score	T03I16	T03I17	T03I18	T03I19	T03I20	T03I21	T03I22
0	44.05%	50.71%	47.07%	27.53%	31.74%	33.49%	51.48%
1	55.56%	49.05%	52.54%	72.09%	67.69%	66.20%	47.85%
O	0.39%	0.24%	0.38%	0.38%	0.56%	0.31%	0.67%

Table 89 Summary of score data for Part 4, crossed, cleaned data set

Score	T04I23	T04I24	T04I25	T04I26	T04I27	T04I28	T04I29	T04I30	T04I31	T04I32	T04I33	T04I34	T04I35
0	59.20%	14.21%	7.62%	26.36%	32.79%	37.12%	16.47%	20.54%	30.38%	21.39%	49.00%	41.89%	15.28%
1	40.25%	85.63%	92.07%	73.22%	66.78%	62.33%	83.06%	78.89%	69.29%	78.31%	50.39%	57.55%	84.22%
O	0.55%	0.17%	0.31%	0.42%	0.44%	0.55%	0.46%	0.56%	0.33%	0.29%	0.61%	0.56%	0.50%

Table 90 Summary of score data for Part 4, sample data set

Score	T04I23	T04I24	T04I25	T04I26	T04I27	T04I28	T04I29	T04I30	T04I31	T04I32	T04I33	T04I34	T04I35
0	56.64%	15.12%	10.07%	27.36%	36.05%	40.06%	21.50%	24.28%	29.19%	25.88%	50.29%	39.66%	17.90%
1	42.64%	84.66%	89.39%	72.04%	63.39%	59.36%	77.87%	74.93%	70.38%	73.77%	48.83%	59.63%	81.36%
O	0.72%	0.22%	0.54%	0.60%	0.56%	0.58%	0.62%	0.78%	0.42%	0.35%	0.88%	0.70%	0.74%

Appendix 7: score distributions

Table 91 Score distributions for each test part

Part	1					Part	2				
Number Correct	Freq-uency	Cumul. Freq.	Percent-age	Cumul. %-age		Number Correct	Freq-uency	Cumul. Freq.	Percent-age	Cumul. %-age	
0	62	62	0.6	0.6	|#	0	25	25	0.3	0.3	|
1	224	286	2.2	2.9	|##	1	131	156	1.3	1.6	|#
2	421	707	4.2	7.1	|####	2	372	528	3.7	5.3	|####
3	881	1588	8.8	15.9	|#########	3	677	1205	6.8	12.1	|#######
4	1161	2749	11.7	27.6	|############	4	1285	2490	12.9	25	|############
5	2571	5320	25.8	53.4	|##########################..	5	1768	4258	17.7	42.7	|##################
6	2076	7396	20.8	74.2	|####################	6	2152	6410	21.6	64.4	|######################
7	2565	9961	25.8	100	|##########################..	7	2050	8460	20.6	84.9	|#####################
						8	1501	9961	15.1	100	|###############
Part	3					Part	4				
Number Correct	Freq-uency	Cumul. Freq.	Percent-age	Cumul. %-age		Number Correct	Freq-uency	Cumul. Freq.	Percent-age	Cumul. %-age	
0	355	355	3.6	3.6	|####	0	7	7	0.1	0.1	|
1	872	1227	8.8	12.3	|#########	1	16	23	0.2	0.2	|
2	1345	2572	13.5	25.8	|##############	2	55	78	0.6	0.8	|#
3	1639	4211	16.5	42.3	|################	3	130	208	1.3	2.1	|#
4	1317	5528	13.2	55.5	|#############	4	279	487	2.8	4.9	|###
5	1719	7247	17.3	72.8	|##################	5	448	935	4.5	9.4	|####
6	402	7649	4	76.8	|####	6	713	1648	7.2	16.5	|#######
7	2312	9961	23.2	100	|#######################	7	973	2621	9.8	26.3	|##########
						8	1299	3920	13	39.4	|#############
						9	1491	5411	15	54.3	|###############
						10	1516	6927	15.2	69.5	|###############
						11	1413	8340	14.2	83.7	|##############
						12	1073	9413	10.8	94.5	|###########
						13	548	9961	5.5	100	|######

Table 92 Score distribution for all test parts together

all Parts					
Number Correct	Freq-uency	Cumul. Freq.	Percent-age	Cumul. %-age	
--------	--------	--------	--------	--------	
0	0	0	0	0	|
1	0	0	0	0	|
2	0	0	0	0	|
3	1	1	0	0	|
4	1	2	0	0	|
5	1	3	0	0	|
6	11	14	0.1	0.1	|
7	8	22	0.1	0.2	|
8	31	53	0.3	0.5	|
9	47	100	0.5	1	|
10	54	154	0.5	1.5	|#
11	79	233	0.8	2.3	|#
12	127	360	1.3	3.6	|#
13	152	512	1.5	5.1	|##
14	185	697	1.9	7	|##
15	236	933	2.4	9.4	|##
16	284	1217	2.9	12.2	|###
17	344	1561	3.5	15.7	|###
18	378	1939	3.8	19.5	|####
19	433	2372	4.3	23.8	|####
20	467	2839	4.7	28.5	|#####
21	506	3345	5.1	33.6	|#####
22	560	3905	5.6	39.2	|######
23	574	4479	5.8	45	|######
24	628	5107	6.3	51.3	|######
25	613	5720	6.2	57.4	|######
26	585	6305	5.9	63.3	|######
27	597	6902	6	69.3	|######
28	558	7460	5.6	74.9	|######
29	530	7990	5.3	80.2	|#####
30	503	8493	5	85.3	|#####
31	449	8942	4.5	89.8	|#####
32	370	9312	3.7	93.5	|####
33	315	9627	3.2	96.6	|###
34	221	9848	2.2	98.9	|##
35	113	9961	1.1	100	|#

Appendix 8: descriptive statistics for candidate background data

Table 93 Most commonly stated candidate L1s

			Original	Sampled	
		2007 data	2005 data	2005 data	
1	**Spanish**	24.00%	41.14%	5.30%	46.44%
2	**Portuguese**	5.00%	9.90%	5.14%	15.04%
3	**French**	5.00%	5.80%	5.35%	11.15%
4	**Swiss-German**	N/A	5.65%	5.19%	10.84%
5	**German**	10.00%	5.00%	4.88%	9.87%
6	**Czech**	3.00%	4.12%	5.04%	9.16%
7	**Italian**	10.00%	3.79%	5.37%	9.16%
8	**Korean**	N/A	2.38%	6.16%	8.54%
9	**Japanese**	N/A	1.74%	4.51%	6.25%
10	**Polish**	7.00%	1.65%	4.27%	5.91%
11	**Greek**	4.00%	0.00%	0.00%	0.00%
12	**Russian**	2.00%	0.34%	0.89%	1.24%
13	**Catalan**	2.00%	0.78%	2.02%	2.80%
	No response	20.00%	11.61%	30.08%	41.69%

Table 94 Candidate age groups

		Original	**Sampled**
	2007 data	**2005 data**	**2005 data**
15 or under	20.00%	10.80%	6.28%
16 - 18	42.00%	38.62%	25.73%
19 - 22	16.00%	20.91%	23.59%
23 - 30	15.00%	19.58%	28.37%
31 or above	6.00%	10.09%	16.02%
No response	1.00%	2.75%	5.96%

Table 95 Candidate gender

		Original	**Sampled**
	2007 data	**2005 data**	**2005 data**
Female	58.00%	58.62%	58.03%
Male	40.00%	37.05%	34.36%

Unknown	N/A	2.40%	2.62%
No response	2.00%	1.93%	4.99%

Table 96 Candidate educational level

		Original	Sampled
Education Level	**2007 data**	**2005 data**	**2005 data**
Primary School	5.00%	0.53%	0.55%
Secondary School	45.00%	42.52%	26.97%
College or University	27.00%	34.62%	41.03%
No response	27.50%	22.33%	31.45%

Table 97 Candidate exam preparation

		Original	Sampled
	2007 data	**2005 data**	**2005 data**
Attended classes	87.00%	88.58%	83.91%
Didn't attend	11.00%	8.11%	9.32%
No response	2.00%	3.31%	6.78%

Appendix 9: descriptive statistics for test materials

Table 98 Descriptive statistics for test materials

	Reference	Combined	2005 test parts			
	Data	**2005**	**Part 1**	**Part 2**	**Part 3**	**Part 4**
Overall number of words	2000	2689	778	734	566	611
Mean words per sentence	18.40	15.11	16.55	14.12	14.90	14.90
Flesch reading ease	66.50	69.84	65.67	73.90	72.43	67.35
Flesch-Kincaid grade level	8.40	7.27	8.21	6.46	6.86	7.57
Tokens	17332	2696	783	734	564	615
Types	3404	1270	335	342	289	304
Type-token ratio	0.20	0.47	0.43	0.47	0.51	0.49
Tokens per type	5.09	2.12	2.34	2.15	1.95	2.02
K1 words	82.24%	82.97%	86.72 %	81.04 %	81.04 %	82.76 %
K2 words	6.65%	4.97%	4.73%	5.46%	5.46%	4.07%
AWL words	3.30%	4.19%	3.32%	1.91%	1.91%	8.78%
Off AWL list words	7.81%	7.87%	5.24%	11.60 %	11.60 %	4.39%
Lexical density	0.5	0.51	0.51	0.51	0.52	0.5

Appendix 10: instructions for selection of relevant text

Coding instructions
Please identify the following:

a) the text you feel a candidate would need in order correctly select the key

b) the text you feel would be likely to make a candidate select a particular distractor (for each distractor).

Please copy the text from the test paper and paste it into the grid cell below. Do not be concerned about formatting.

It may be that in some cases you cannot identify specific text for an item. If so, please state this in the relevant cell.

Thank you for your help.

part 1

	a	b	c	d	e	f	g	h	i
1									
2									
3									
4									
5									
6									
7									

part 2

	a	b	c	d
8				
9				
10				
11				
12				
13				
14				
15				

part 3

	a	b	c	d	e	f	g	h	i
16									
17									
18									
19									
20									
21									
22									

part 4

	a	b	c	d
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				

Appendix 11: incidence matrix summary

Figures in the tables in this section are the mean for continuous-based indicators, and the mode for factor-based indicators.

Table 99 Incidence matrix summarised by test part, OP component

Part	X001.OP.syll	X002.OP.BNC	X002.OP.BNC.CLPS	X003.OP.AWL	X006.OP.CELEX.cont.f	X007.OP.CELEX.all.f.log	X008.OP.CELEX.cont.log	X009.OP.type.tok	X010.OP.hypernymy	X011.OP.polysemy	X012.OP.lex.density	X013.OP.concrete	X014.OP.mod.noun	X015.OP.left.emb	X016.OP.neg	X017.OP.hol.neg	X017.OP.hol.neg.CLPS	X018.OP.fronted	X019.OP.passive	X020.OP.connect	X021.OP.stem.o	X022.OP.props	X000.OP.prop.dens	X023.OP.causal	X024.OP.intent	X025.OP.temp
1	1.57	14	14	15	2.39	3	0.98	0.99	1.77	4.96	40.3	340	0.45	1.88	0	6	6	0	15.63	51.57	0	63	0.56	0.19	0.25	-2
2	1.44	8	8	4	2.3	3.04	1.09	0.99	1.67	5.41	31	367	0.67	1.18	6.82	2	2	0	15.53	61.01	0.09	35.1	0.53	0.2	0.44	0.55
3	1.5	26	26	4	2.28	3.06	1.31	0.99	1.7	3.94	28.5	387	0.9	3.5	11.4	1	1	1	0	36.2	0	84	0.52	0	0.13	-1.3
4	1.75	3	1	3	2.38	2.76	1.01	1	2.7	5	65.4	443	0.97	2	0	0	1	0	0	12.5	0	6.77	0.5	0	0.04	-2

Table 100 Incidence matrix summarised by test part, SEARCH component

Part	X051.SEARCH.order	X052.SEARCH.demarc	X052.SEARCH.demarc.CLPS	X053.SEARCH.LSA.term
1	1	2	2	0.82
2	1	0	1	0.77
3	1	1	1	0.75
4	0	0	1	0.71

Table 101 Incidence matrix summarised by test part, READ component, first 15 indicators

X026.READ.syll	X027.READ.BNC	X027.READ.BNC.CLPS	X028.READ.AWL	X031.READ.CELEX.cont.f	X032.READ.CELEX.all.f.log	X033.READ.CELEX.cont.log	X034.READ.type.tok	X035.READ.hypernymy	X036.READ.polysemy	X037.READ.lex.density	X038.READ.concrete	X039.READ.mod.noun	X040.READ.left.emb	X041.READ.neg
1.48	16	1	4	2.34	2.97	0.61	0.97	1.5	4.05	14	363	0.65	4.19	3.03
1.38	16	3	4	2.36	3	1.33	0.99	1.49	4.42	16.6	372	0.87	2.76	2.93
1.41	26	3	4	2.36	3.01	1.27	0.95	1.77	4.37	14.9	369	0.83	2.87	2.25
1.56	16	3	4	2.29	3.01	0.84	0.99	1.79	4.82	15.8	385	0.8	3.42	0.36

Table 102 Incidence matrix summarised by test part, READ component, last 14 indicators

Part	X042.READ.hol.neg	X042.READ.hol.neg.CLPS	X043.READ.fronted	X043.READ.fronted.CLPS	X044.READ.passive	X045.READ.connect	X046.READ.stem.o	X047.READ.props	X047.READ.props.CLPS	000.READ.prop.dens	X048.READ.causal	X049.READ.intent	X050.READ.temp	X000.READ.sentence
1	8	1	0	0	8.03	91.4	0.04	128	3	0.56	0.18	0.34	-0.8	13.1
2	0	1	0	0	1.34	101	0	75.6	2	0.53	0.31	0.23	0.44	11.5
3	0	0	0	0	2.71	81	0.13	103	2	0.52	0.47	0.85	0.31	12.6
4	0	0	1	1	4.74	117	0.11	58.5	2	0.5	0.65	0.94	-0.5	5.62

Table 103 Incidence matrix summarised by test part, RD component

Part	X054.RD.LSA.term.KEY	X055.RD.LSA.term.DIST	X056.RD.LSA.doc.KEY	X057.RD.LSA.doc.DIST	X058.RD.disperse	X059.RD.pract
1	0.83	0.78	0.42	0.13	84.3	2
2	0.81	0.82	0.31	0.13	73.6	1.97
3	0.77	0.75	0.19	0.01	59	2.88
4	0.65	0.66	0.12	-0.1	118	2.24

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing* (7th ed.). Washington, D.C.: AERA.

Albano, A. D. (2014). Equate: observed-score linking and equating (Version R package version 2.0-3). Retrieved from http://CRAN.R-project.org/package=equate

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2010). "Cognitive Diagnosis and Q-Matrices in Language Assessment": A Commentary. *Language Assessment Quarterly, 7*(1), 96-103.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing, 30*(4), 535-556.

ALTE, & Council of Europe. (2011). *Manual for Language Test Development and Examining for use with the CEFR*. Retrieved from http://www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf

Andrich, D., & Kreiner, S. (2010). Quantifying Response Dependence Between Two Dichotomous Items Using the Rasch Model. *Applied Psychological Measurement, 34*(3), 181-192.

Aryadoust, V., & Goh, C. C. M. (2014). Predicting Listening Item Difficulty with Language Complexity Measures: a Comparative Data Mining Study. *CaMLA Working Papers* (Vol. 2). Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/CWP-2014-02.pdf

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and context in defining constructs in language assessment. In J. Fox & M. Wesche (Eds.), *Language testing reconsidered* (pp. 41-71).Ottawa: University of Ottawa Press.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language - the Cambridge-TOEFL Comparability Study*. Cambridge: Cambridge University Press.

Bachman, L. F., & Eignor, D. R. (1997). Recent advances in quantitative test analysis. In C. Clapham & D. Corson (Eds.), *Language Testing and Assessment* (Vol. 7) (pp. 227-242). Dordrecht: Kluwer.

Bachman, L. F., & Palmer, A. S. (1982). The Construct Validation of Some Components of Communicative Proficiency. *TESOL Quarterly, 16*, 449.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment In Practice: Developing Language Assessments And Justifying Their Use In The Real World*. Oxford: Oxford University Press.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-5). Retrieved from http://lme4.r-forge.r-project.org/

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441-165.

Bejar, I. I. (2010). Recent Development and Prospects in Item Generation. In S. E. Embretson (Ed.), *Measuring psychological constructs: advances in model-based approaches* (1st ed., pp. 201-226). Washington, DC: American Psychological Association.

BNC Consortium. (2007). The British National Corpus (Version 3). Oxford: University of Oxford Computing Services. Retrieved from http://www.natcorp.ox.ac.uk/

Bolt, D. M., & Lall, V. F. (2003). Estimation of Compensatory and Noncompensatory Multidimensional Item Response Models Using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.

Brown, C., Snodgrass, T., Covington, M. A., Han, J., Boisclair, C., Drucker, E., . . . Nadig, K. (2012). CPIDR 5.1 Computerized Propositional Idea Density Rater. Athens, GA: University of Georgia Research Foundation. Retrieved from http://www.ai.uga.edu/caspr

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods, 40*(2), 540-545.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3), 423-466.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Canale, M., & Swain, M. (1980). Theoretical basis of communicative approaches to second language teaching and testing. *Applied Linguistics, I*(1), 1-47.

Carr, T. H., Brown, T. L., Vavrus, L. G., & Evans, M. A. (1990). Cognitive Skill Maps and Cognitive Skill Profiles: Componential Analysis of Individual Differences in Children's Reading Efficiency. In T. H. Carr & B. A. Levy (Eds.), *Reading and Its Development : Component Skills Approaches*. San Diego, CA: Academic Press Inc.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller (Ed.), *Issues in language testing* (pp. 80-105). Rowley, MA: Newbury House.

Castello, E. (2008). *Text Complexity and Reading Comprehension Tests*. Pieterlen: Peter Lang.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1).

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*(4), 369-383.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge Univerity Press.

Child, D. (2006). *The Essentials of Factor Analysis* (3rd ed.). London: Continuum International.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: Massachusetts Institute of Technology Press.

Chou, C.-P., & Huh, J. (2012). Model Modification in Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 232-246). New York, NY: Guilford Press.

Cobb, T. (2013). VocabProfile v4. 31/03/13, from http://www.lextutor.ca/vp/

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A, 33*(4), 497-505.
Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
Covington, M. A. (2012). *CPIDR® 5.1 USER MANUAL* Retrieved from http://www.ai.uga.edu/caspr/CPIDR-5-Manual.pdf
Coxhead, A. (1998). *An Academic Word List*. Wellington: Victoria University of Wellington.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.
De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: NY: Guilford Publications, Incorporated.
De Boeck, P., Bakker, M., Zwister, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of Item Response Theory models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1-28.
De Boeck, P., & Wilson, M. (2004a). Descriptive and explanatory item response models. In P. de Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 43-74). New York, NY: Springer.
De Boeck, P., & Wilson, M. (2004b). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*: Springer.
DeMars, C. E. (2010). *Item Response Theory*. Oxford: Oxford University Press.
Dennis, S. (2011). How to use the LSA website. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57-70). Abingdon: Routledge.
Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: with the lme4 package. *Journal of Statistical Software, 20*(2), 1-18.
Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.
Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175-186.
Embretson, S. E. (1985). Studying Intelligence with Test Theory Models. In D. K. Detterman (Ed.), *Current Topics in Human Intelligence* (pp. 98-140). Norwood, NJ: Ablex Publishing Company.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Embretson, S. E., & Wetzel, C. D. (1987). Component Latent Trait Models for Paragraph Comprehension Tests. *Applied Psychological Measurement, 11*(2), 175-193.
Embretson, S. E., & Yang, X. (2006). Multicomponent Latent Trait Models for Complex Tasks. *Journal of Applied Measurement, 7*(3), 335-350.
Embretson, S. E., & Yang, X. (2013). A Multicomponent Latent Trait Model for Diagnosis. *Psychometrika, 78*(1), 14-36.
Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 11-28). Rowley, MA: Newbury House.
Fellbaum, C. (1998). WordNet - An Electronic Lexical Database. Cambridge, MA: MIT Press.

Field, J. (2013). Cognitive Validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining Listening: Research and Practice in Assessing Second Language Listening*: Cambridge University Press.
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359-374.
Fischer, G. H. (1995). The Linear Logistic Test Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications* (pp. 131-155). New York, NY: Springer.
Fouly, K. A., Bachman, L. F., & Cziko, G. A. (1990). The Divisibility of Language Competence: A Confirmatory Approach. *Language Learning, 40*(1), 1-21.
Fox, J. (2002). *An R and S-Plus Companion to Applied Regression*. Thousand Oaks, CA: SAGE Publications.
Freedle, R., & Kostin, I. (1993). The Prediction of TOEFL Reading Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items *TOEFL Research Reports*. Princeton, NJ: ETS.
Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28*(1), 77-104.
Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
Geranpayeh, A., & Taylor, L. (2013). *Examining Listening: Research and Practice in Assessing Second Language Listening* (Vol. 35). Cambridge: Cambridge University Press.
Gorin, J. S. (2005). Manipulating Processing Difficulty of Reading Comprehension Questions: The Feasibility of Verbal Item Generation. *Journal of Educational Measurement, 42*(4), 351-373.
Gorin, J. S., & Embretson, S. E. (2006). Item Diffficulty Modeling of Paragraph Comprehension Items. *Applied Psychological Measurement, 30*(5), 394-411.
Gorin, J. S., & Svetina, D. (2012). Cognitive Psychometric Models as a Tool for Reading Assessment Engineering. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Reaching an Understanding: Innovations in How We View Reading Assessment* (pp. 169-184). Lanham, MD: R&L Education.
Grabe, W. (2009). *Reading in a Second Language: Moving from Theory to Practice.* Cambridge: Cambridge University Press.
Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher, 40*(5), 223-234.
Gwet, K. L. (2012). Handbook of Inter-rater Reliability. Gaithersburg, MD: Advanced Analytics, LLC.
Halliday, M. A. K., & Hasan, R. (1990). *Cohesion in English* (1 ed.). Harlow: Longman.
Hawkey, R. (2009). *Examining FCE and CAE*. Cambridge: Cambridge University Press.
Hawkins, J. A., & Buttery, P. (2012). *Criterial Features in L2 English - Specifying the Reference Levels of the Common European Framework* (Vol. 1). Cambridge: Cambridge University Press.
Hulstijn, J. H. (2011). Language Proficiency in Native and Nonnative Speakers: An Agenda for Research and Suggestions for Second-Language Assessment. *Language Assessment Quarterly, 8*(3), 229-249.
Hulstijn, J. H. (2014). The Common European Framework of Reference for Languages: A challenge for applied linguistics. *ITL - International Journal of Applied Linguistics, 165*(1), 3-18.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.
IBM Corp. (2013). IBM SPSS Statistics for Windows (Version 22.0). Armonk, NY: IBM Corp.
Ilc, G., & Stopar, A. (2014). Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English. *Language Testing – Online First*. Retrieved from http://ltj.sagepub.com/content/early/2014/12/17/0265532214562098.full.pdf+html
Jackson, N. E. (2005). Are university students' component reading skills related to their text comprehension and academic achievement? *Learning and Individual Differences, 15*(2), 113-139.
Jang, E. E. (2009). Demystifying a Q-Matrix for Making Diagnostic Inferences About L2 Reading Skills. *Language Assessment Quarterly, 6*(3), 210-238.
Jones, N. (1998). Classic (Version 1.00). Cambridge: Constructs Learning & Assessment Ltd.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.
Kane, M. T. (2009). Validating the Interpretations and Uses of Test Scores. In R. W. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications* (pp. 39-64). Charlotte, NC: Information Age Publishing.
Kane, M. T. (2011). The Errors of Our Ways. *Journal of Educational Measurement, 48*(1), 12-30.
Kenny, D. A., & Milan, S. (2012). Identification - A nontechnical discussion of a technical issue. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 145-163). New York, NY: Guilford Press.
Khalifa, H., & Weir, C. J. (2009). *Examining Reading: research and practice in assessing second language reading* (Vol. 29). Cambridge: Cambridge University Press.
Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.
Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.
Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring Document Literacy: Variables Underlying the Performance of Young Adults. *Reading Research Quarterly, 25*(1), 5-30.
Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in Psychometric Research. *Educational and Psychological Measurement, 69*(2), 232-244.
Laham, D. (1998). Latent Semantic Analysis Website. from http://lsa.colorado.edu/
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.
Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2011). *Handbook of Latent Semantic Analysis*. Abingdon: Routledge.
Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing, 21*(1), 74-100.
Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York, NY: Cambridge University Press.
McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2005). Coh-Metrix version 1.4. from Institute for Intelligent Systems, University of Memphis http//:cohmetrix.memphis.edu
McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2012). Coh-Metrix version 3.0. Memphis, TN: Department of Psychology, University of Memphis. Retrieved from http//:cohmetrix.memphis.edu.
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (Third ed., pp. 13-103). New York, NY: ACE/Macmillan.
Microsoft Corp. (2010a). Microsoft Excel 2010 (Version 2010). Redmond WA: Microsoft.
Microsoft Corp. (2010b). Microsoft Word 2010 (Version 2010). Redmond WA: Microsoft.
Mislevy, R. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.
Oller, J. W. (1983). Evidence for a general language proficiency factor: an expectancy grammar. In J. W. Oller (Ed.), *Issues in Language Testing Research* (pp. 3-10). Rowley, MA: Newbury House.
Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). CONCRETENESS, IMAGERY, AND MEANINGFULNESS VALUES FOR 925 NOUNS. *Journal of Experimental Psychology, 76*(1, Pt.2), 1-25.
R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org
Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (New expanded ed. ed.). Chicago, IL: University of Chicago Press.
Reckase, M. D. (1994). What is the 'correct' dimensionality for a set of item response data? In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern Theories of Measurement: Problems and Issues* (Kindle ed., pp. 87-92). Ottawa: Edumetrics Research Group, University of Ottawa.
Reckase, M. D. (2009). *Multidimensional Item Response Theory* (Kindle ed.). New York, NY: Springer.
Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research, 47*(5), 667-696.
Rijmen, F., & De Boeck, P. (2002). The Random Weights Linear Logistic Test Model. *Applied Psychological Measurement, 26*(3), 271-285.
Rouet, J.-F. (2003). What was I looking for? The influence of task specificity and prior knowledge on students' search strategies in hypertext. *Interacting with Computers, 15*(3), 409-428.
Rouet, J.-F. (2012). *The Skills of Document Use: From Text Comprehension to Web-Based Learning*. New York, NY: Routledge.
Rouet, J.-F., Vidal-Abarca, E., Erboul, A. B., & Millogo, V. (2001). Effects of Information Search Tasks on the Comprehension of Instructional Text. *Discourse Processes, 31*(2), 163-186.
Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing, 23*(4), 441-474.
Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining Multiple Regression and CART to Understand Difficulty in Second Language Reading and Listening Comprehension Test Items. *International Journal of Testing, 1*(3/4), 185.
SAS Institute Inc. (SAS). SAS. Cary NC: SAS Institute Inc.

Saville, N. (2010). *Setting quality standards in international language assessment*. Paper presented at the 39th ALTE Meeting Conference Day, Prague.

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-Matrix Construction: Defining the Link Between Constructs and Test Items in Large-Scale Reading and Listening Comprehension Assessments. *Language Assessment Quarterly, 6*(3), 190-209.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5-30.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: research and practice in assessing second language writing* (Vol. 26). Cambridge: Cambridge University Press.

Sheehan, K. M., & Ginther, A. (2000). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section*. Paper presented at the Annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners.* (Vol. 32). Cambridge: Cambridge University Press.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99-128.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(1), 25-40.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing, 25*(4), 435-464.

Sternberg, R. J. (1985). Componential Analysis: a Recipe. In D. K. Detterman (Ed.), *Current Topics in Human Intelligence* (pp. 179-201). Norwood, NJ: Ablex Publishing Company.

Stroup, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Boca Raton, FL: Taylor & Francis.

Taylor, L. (2014). General Language Proficiency (GLP): Reflections on the "Issues Revisited" from the Perspective of a UK Examination Board. *Language Assessment Quarterly, 11*(2), 136-151.

Taylor, L. (Ed.). (2011). *Examining Speaking - Research and practice in assessing second language speaking* (Vol. 30). Cambridge: Cambridge University Press.

Thissen, D., & Steinberg, L. (2010). Using Item Response Theory to Disentangle Constructs at Different Levels of Generality. In S. E. Embretson (Ed.), *Measuring psychological constructs: advances in model-based approaches* (pp. 123-144). Washington, DC: American Psychological Association.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. de Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer.

University of Cambridge ESOL Examinations. (2007). *First Certificate English - Handbook for Teachers for Examinations from December 2008*. Cambridge: University of Cambridge ESOL Examinations.

University of Cambridge ESOL Examinations. (2008). *Cambridge English Advanced - Handbook for Teachers*. Cambridge: University of Cambridge ESOL Examinations.

University of Memphis. (2012). Coh-Metrix version 3.0 indices. from http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a Second Language: Process, Product, and Practice*. London: Prentice Hall.
van der Linden, W. J. (2005). *Linear Models for Optimal Test Design* (1 ed.). New York, NY: Springer.
van Steensel, R., Oostdam, R., & van Gelderen, A. (2013). Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language Testing, 30*(1), 3-21.
Vollmer, H. J., & Sang, F. (1983). Competing hypotheses about second language ability: a plea for caution. In J. W. Oller (Ed.), *Issues in language testing* (pp. 29-74). Rowley, MA: Newbury House.
Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge: Cambridge University Press.
Wald, A., & Wolfowitz, J. (1940). On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics, 11*(2), 147-162.
Weir, C. J. (2005). *Language testing and validation: an evidence-based approach.* Basingstoke: Palgrave Macmillan.
Weir, C. J. (2011). *Context Validity: a quantitative approach.* Paper presented at the ALTE Meeting, Bochum.
Weir, C. J. (2013). Appendix B - Case study: a quantitative analysis of context validity of the CPE reading passages used in translation tasks (1913-88), summary tasks (1930-2010) and comprehension qustion (MCQ/SAQ) tasks (1940-2010). In C. J. Weir, I. Vidaković, & E. D. Galaczi (Eds.), *Measured constructs: A history of Cambridge English language examinations 1913-2012* (pp. 472-537). Cambridge: Cambridge University Press.
Weir, C. J., Hawkey, R., Green, A., & Devi, S. (2012). The cognitive processes underlying the academic reading constructs as measured by IELTS. In C. Weir & L. Taylor (Eds.), *IELTS collected papers 2: research in reading and listening assessment* (Vol. 34) (pp. 212-269). Cambridge: Cambridge University Press.
Weir, C. J., Hughes, A., & Porter, D. (1990). Reading skills: hierarchies, implicational relationships and identifiability. *Reading in a Foreign Language, 7*(1), 505-510.
West, M. (1953). *A General Service List of English Words*. London: Longman.
Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479-494.
Wickham, H. (2013). reshape: Flexibly reshape data (Version 0.84). Retrieved from http://had.co.nz/reshape
Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach* (1 ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
Wilson, M., & Moore, S. (2011). Building out a measurement model to incorporate complexities of testing in the language domain. *Language Testing, 28*(4), 441-462.
Wilson, M., & Moore, S. (2012). An Explanative Modeling Approach to Measurement of Reading Comprehension. In J. P. Sabatini, T. O'Reilly, & E. R. Albro (Eds.), *Reaching an Understading - Innovations in How We View Reading Assessment* (pp. 147-168). Lanham, MD: Rowman & Littlefield Education.
Wu, R. Y. (2014). *Validating Second Language Reading Examinations* (Vol. 41). Cambridge: Cambridge University Press.
Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8*(2), 125-145.

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing, 25*(4), 521-551.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162-185.

Zwaan, R. A., & Singer, M. (2003). Text Comprehension. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of Discourse Processes* (pp. 83-121). Mahwah, NJ: Lawrence Erlbaum.